# Measuring the Evolution of Innovation in Patent Citation Networks

**Ryan Steed***
Departments of Computer Science & Economics
The George Washington University
Washington, D.C. 20052
ryansteed@gwu.edu

November 10, 2019

**Abstract**

This investigation measures the propagation of innovation and the evolution of knowledge by examining the changing structures of patent citation networks. What is the effect of government-sponsored innovation and intellectual property law on the transfer of knowledge in the patent citation network network? An index of total knowledge contribution (TKC) is developed to to measure the impact of the intellectual property in individual patents on subsequent inventions. The index is applied to citation networks constructed from patents granted between 1976 and 2018 in a variety of USPC technology sectors. A simple OLS regression finds that government patents do tend to contribute more, but the application of a time series model and a pooled OLS regression shows that the recent America Invents Act significantly reduced patent contribution.

***Keywords*** Innovation · Knowledge transfer · Patent citation network · Patent importance · America Invents Act

## 1 Introduction

Economists, historians and business leaders generally agree that innovation is inextricably linked to continued prosperity and national competitiveness. Accordingly, nations sponsor research and craft legislation, such as intellectual property protection, to stimulate innovation. Yet, to persuade a public skeptical of government expenditures, leaders and policymakers often seek ways to assess this investment and quantify its benefits. This study aims to address the broad practical question: is there a rigorous way to quantifiably assess innovation and its spread with currently available data?

United States patent filings, often the first public release of information about cutting-edge industry technology, comprise an ever-growing complex network of documents linked by citations and rigid U.S. Patent & Trademark Office (USPTO) classifications. Unlike text-based networks, citation network edges are consistently defined and generally regarded as useful structural representations of patent documentation. Open source patent data can be used to construct a direct citation network that depicts a complex landscape across industries useful for a variety of technology, inventor, and firm analysis (Henrique et al., 2018). Oh et al. (2017) establish base network statistics for measuring technological innovation over time, including degree centrality and the Hirsch index (H-index), a measure of individual or firm productivity used in social and financial networks. Lanjouw and Schankerman (2004) define a custom patent quality index using four separate observable characteristics of patent documents. More complex models attempt leverage citation network structure for forecasting. Chang et al. (2015) use the H-index as an indicator of the correlation between patent citation count and market value for individual firms, linking moderately productive firms to high market value. Meanwhile, Daim et al. (2006) construct a system dynamics model that uses bibliometric techniques to identify emerging technologies in three separate industries. Other studies have attempted to link patent data as such to market value, breakthrough technologies, and even GDP (Karanikić et al., 2017; Lanjouw and Schankerman, 2004).

*The author acknowledges the sponsorship of the GW Data MASTER Program, supported by the National Science Foundation under grant DMS-1406984.

The size and complexity of the patent citation network make it ideal for network evolution analysis, from large-network generalizations to case-specific claims (Leskovec et al., 2007). As of 1999, the direct citation network for publicly released patent fillings from the USPTO database contained nearly 4 million nodes and over 16.5 million edges and is nearly connected (Leskovec et al., 2007). Martinelli and Nomaler (2014) cast concepts of genetic inheritance to patent relations, identifying interruptions in patent development over time by breaks in the "lineage" of a particular subfield. Most pertinent is the work of You et al. (2017), who apply two time series models (Bass and ARIMA) to a single USPTO patent subclass to forecast a customized indicator of "development potential."

However, little investigation has been conducted into specific policy problems. For instance, the 2011 America Invents Act (AIA) disrupted the United States patenting process, replacing first-to-invent policy with the first-to-file system popular in Europe and sparking a furious debate over the impact of the policy on innovation and technological growth. The institution of patent protection, an integral part of intellectual property law and economic competition, is fundamentally changed by its passage, but many of its effects are unknown and unexplored. Miyagiwa (2015) conducts a simple patent race model to determine the effects of the AIA on innovation, determining by qualitative criteria that first-to-file is a slightly worse catalyst for innovation. Pierce (2012), who considers it "the most substantial change to American patent law since the Patent Act of 1952," holds that legal stipulations in the Act will inhibit collaborative research (perhaps resulting in a citation network with higher modularity). Research on the network effects of the AIA is lacking, and nearly no temporal analyses of the policy change exist.

To understand the progress of innovation in competitive markets, it is essential to examine the legal institutions governing intellectual property; those legal institutions dictate the private research and development practices of technology firms. Further, if patent filing has an effect on the evolution of invention, patent policy plays an extremely important role in technological evolution. Fortunately, public patent data provide the means to track this progression of knowledge and its interaction with government policy.

This investigation attempts to estimate the all-time knowledge impact of any given patent on future inventions by combining existing measures of patent importance. With this metric, we examine the evolution of patent knowledge transfer over time with respect to government support and legislation, identifying the predictors of knowledge contribution and trends in knowledge impact over time.

Section 2 describes patent data collection and citation network construction methods. Section 3 explicates the metrics used to evaluate patent importance, knowledge transfer, and total contribution, and provides descriptive statistics for the data. Finally, Section 5 applies our theoretical framework to elucidate the evolutionary differences between technology sectors and the effects of the AIA on patent knowledge transfer. Section 6 concludes.

## 2 Description of Data

The patent data used for this investigation include U.S. patents granted between January 1976 and November 2018 accessed using the USPTO PatentsView API. Patent data were sampled from several test categories: first, the 21,834 patents in the coherent light generator patent class (U.S. Patent Classification Mainclass 372) used by You et al. (2017) as a targeted small case study in specific industry; second, a random set of the National Bureau of Economic Research (NBER) technology categories (Hall et al., 2001). These data also include several observable characteristics, including features like the number of claims, the time to process, and various industry categories assigned by expert reviewers at the USPTO and NBER, including the Cooperative Patent Classifications and NBER Technology Categories, respectively.

We construct a citation network for each test set using the forward citations of each patent in the set. The citation network is defined as follows: each edge represents the relationship $R$ between patents $x$ and $y$ such that $xRy$ implies that patent $y$ is a forward citation of patent $x$. Each node in the network represents the information contained in a single patent, and edges represent the attribution of knowledge from one information set to another. This characterization of the patent citation relation is based in the legal concept of *prior art*. The citing patent indicates that it builds upon the technological progress of the cited patent, or at least claims that the content of the cited patent is relevant to the originality of the citing patent's claims (Strandburg et al., 2009). Figure 1 displays a sample citation network from the coherent light generator test set.

## 3 Theoretical Framework: Measuring Knowledge Impact

To measure the effect of government policy on the evolution of innovation, we must define and quantify innovation as a mathematical property of a given state of the patent citation network. In this investigation, we focus on innovation as knowledge production. Given the definition of the citation relation posited in this paper, the relations of patents in the

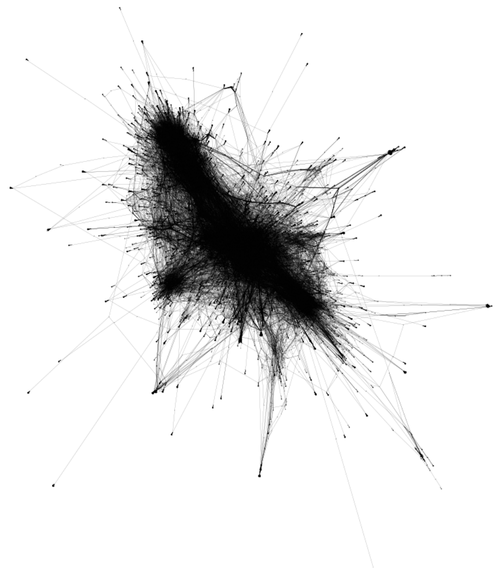

Figure 1: Patent citation network constructed from 5,000 random patents in the coherent light generator mainclass. Note that only connected nodes are included.

citation network indicate a time-based, approximate technological progression. Knowledge flows forward or laterally from the cited to the citing patent, but never backwards. As such, mathematical measures of the topological position of nodes in the directed citation network signal the knowledge reception and contribution of patents and technology groups. In the following section, we will leverage this interpretation of the citation network to fully describe the contribution of any given patent to its descendants, public knowledge, and global innovation.

To avoid conceiving a precise definition of the citation relation, most literature defines "patent importance" as the strength or meaningfulness of a patent's position in the citation network as a loose model of knowledge flow. There are several existing topological measures of patent importance. Most commonly, a patent's importance is simply defined as the number of forward (reverse) citations a patent receives. Another common metric is h-index, the maximum value $h$ such that at least $h$ descendants have at least $h$ forward citations. If $f(i)$ is the number of citations for the patent $i$:

$$h_i = max_i[min(f(i), i)]$$

### 3.1 Forward Importance

The primary motivation for the evolutionary importance metric used in this investigation comes from Trajtenberg et al. (1997), who define a "forward-looking" measurement of patent importance (designated $F$) based on the forward citations of the direct descendants of any patent $i$. They also suggest an unspecified weight $\lambda$ to reduce the impact of descendant importance. Suppose the patent $i$ has $n_i$ descendants, and $f_{i,j}$ is the forward citation count for the $j$th descendant.

$$F_i = n_i + \lambda \sum_{j=1}^{n_i} f_{i,j}$$

While Trajtenberg et al. (1997) evaluate only direct descendants, their measure can easily be calculated recursively to account for *all* descendants. For practical purposes, this recurrence relation may only be computed up to a certain distance from $i$. Recalling that a patent $i$ has descendants $\{1, 2, \ldots, n\}$:

$$F_i^R = n_i + \lambda \sum_{j=1}^{n_i} F_j^R$$

The forward importance method is ideally suited for evaluating knowledge contribution because it judges a patent based on the accomplishments of its descendants. A patent may have few descendants, but if one of its descendants is cited many times, that patent will receive a higher importance rating than it would under traditional rating schemes.

However, the forward importance metric assumes that there is a uniformly proportional relationship between the contribution of the parent patent and the importance of its descendants, which may not be the case. A successful dependent with no other parents is a better indicator of importance than a successful descendant with many parents.

### 3.2 Persistence

Martinelli and Nomaler (2014) separately implement a "persistence" metric (designated $P$) that addresses the issue of descendant knowledge attribution. The persistence index draws inspiration from Mendelian inheritance to create a weighting method based on the number of descendants of the root patent at each depth in the descendant tree. Suppose $l_i$ is the number of nodes in the descendant tree at the same distance from the root patent $r$ as patent $i$. Let $b_{i,j}$ be the backward citation count for the $j$th descendant of $i$. Thus the persistence index includes a "backward-looking" factor to its forward-looking consideration of descendants.

$$P_i = \sum_{j=1}^{l_i} \frac{1}{b_{i,j}}$$

### 3.3 Total Knowledge Contribution

We propose a metric that combines these two approaches - recursive examination of patent descendants and their relative importance persistence indexing in a genetic fashion - with existing topological importance metrics to provide a more robust representation of a patent's contribution to future inventions. Traditional importance metrics deal only with the immediately proximal (forward citations) or global topology (network centrality) of a patent. The forward importance approach begins to describe the contribution of a patent in terms of the contributions of its descendants, but stops after only one generation and does not account for the dwindling influence of ancestor patents as generations of citations go by. The genetic decomposition approach accounts for these factors but does nothing to evaluate the relative importance of patents in the descendant tree, using only nominal, equally weighted proportions.

Our total knowledge contribution index (TKC) combines these two approaches. We evaluate the relative contributions of a patent's descendants with the forward importance method while accounting for how much of each descendant's "heritage" can be attributed to the root patent under investigation with the persistence index. Further, instead of using a simple forward citation measure, the TKC can be modified to use any standard measure of importance. In this investigation, we calculate versions of TKC weighted by the following importance metrics: forward citation count (out-degree centrality) and h-index.

Where $W$ is any existing topological, roughly proximal measure of patent importance (e.g. the h-index, or the out-degree centrality), total patent knowledge impact $K$ is as follows, where the contributions of each successive generation increase total knowledge contribution by a discount factor $0 < \lambda \leq 1$:

$$K_i = W_i + \sum_{j=1}^{n_i} \lambda P_j K_j$$

Note that the function $P_i$ is the persistence index with respect to the root node being evaluated.

$$P_i = \begin{cases} 1 & i = r \\ \frac{1}{b_i} & i \neq r \end{cases}$$

To compute the knowledge metric for a given patent, we construct an individual tree-based patent citation network with the patent as its node. These networks become exponentially large as the depth of the citation search increases; Figure 2 shows the sprawling three-generation descendant tree of patent for an X-ray generator granted in 1976. This patent gave rise to two clusters of descendants, one more prolific than the other.

To see the computation in practice, take the simple citation network considered by Martinelli and Nomaler (2014) in their illustration of the persistence index (Figure 3). Assume patent $I$ and patent $H$ each have two descendants with no forward citations. As an example, consider the knowledge contribution of patent $A$. If there is no discount between generations ($\lambda = 0$) and the forward citation count is used as the weighting key $W$, $K_A$ expands as follows:

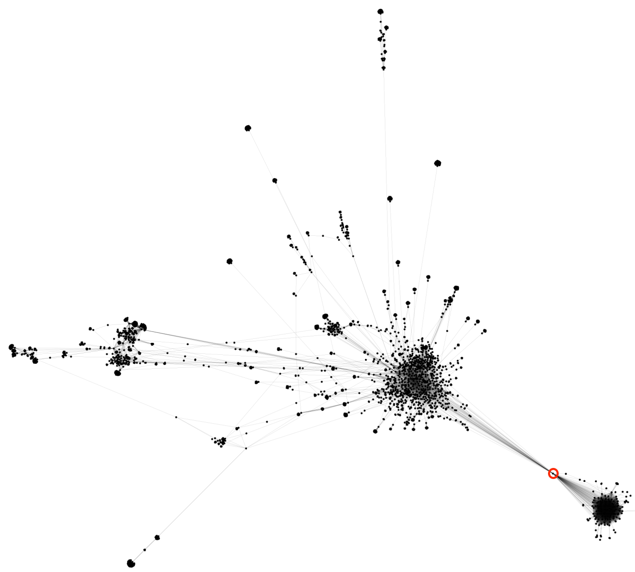

Figure 2: Forward citation network rooted on patent 3961197 (indicated in red) in the coherent light generator class up to three generations. Descendant clusters are displayed with a force-directed layout.

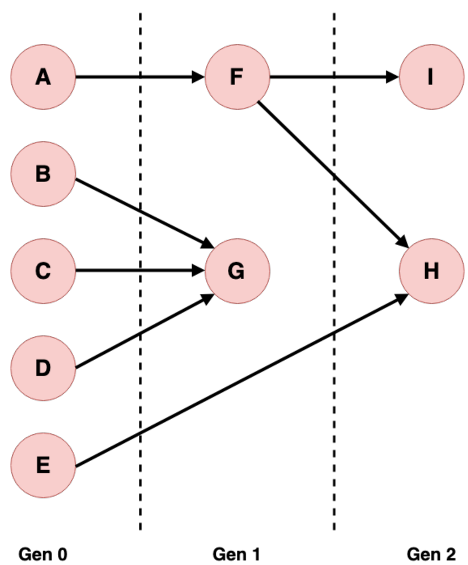


Figure 3: Example citation network.

$$
\begin{aligned}
K_A &= W_A + P_F K_F \\
&= W_A + P_F(W_F + P_I K_I + P_H K_H) \\
&= W_A + P_F(W_F + P_I W_I + P_H W_H)
\end{aligned}
$$

The patent rankings produced by the TKC are fairly intuitive. $G$ is never cited and receives a score of zero. $B$, $C$, and $D$ are all cited by $G$, and so receive minimal scores. $E$ is the sole citation of $H$, which in turn has two children, and

| $i$ | $P_i$ |
|---|---|
| F | 1 |
| G | $\frac{1}{3}$ |
| I | 1 |
| H | $\frac{1}{2}$ |

| $i$ | $K_i$ |
|---|---|
| A | 6 |
| B | 1 |
| C | 1 |
| D | 1 |
| E | 2 |
| F | 5 |
| G | 0 |

Figure 4: Example knowledge impact computation for the example network in Figure 3, with no discount.

receives a modest score. Patent $F$ is the sole citation of patent $I$ and co-contributes to the knowledge in patent $H$, and receives a high score. So does its parent, $A$.

$A$ and $F$ clearly contribute the most to posterity. When $\lambda = 1$, $A$ is credited for all of the contributions of $F$ and has the maximum contribution. But for most discount rates less than one, $F$ has the higher TKC index - a lower discount rate assumes that child patents contain more original content than a higher discount rate. Figure 5 shows the relationship between TKC and the discount rate $\lambda$ for this example. Note that the TKC index for patent $E$ does not exceed that of $A$ or $F$ for any value of the discount constant; this result also makes intuitive sense, since the patent $E$ is an equal contributor to $H$ and does not contribute to any other patent. This investigation will hereafter use a discount rate of 0.5 according to Trajtenberg et al. (1997)'s justification.

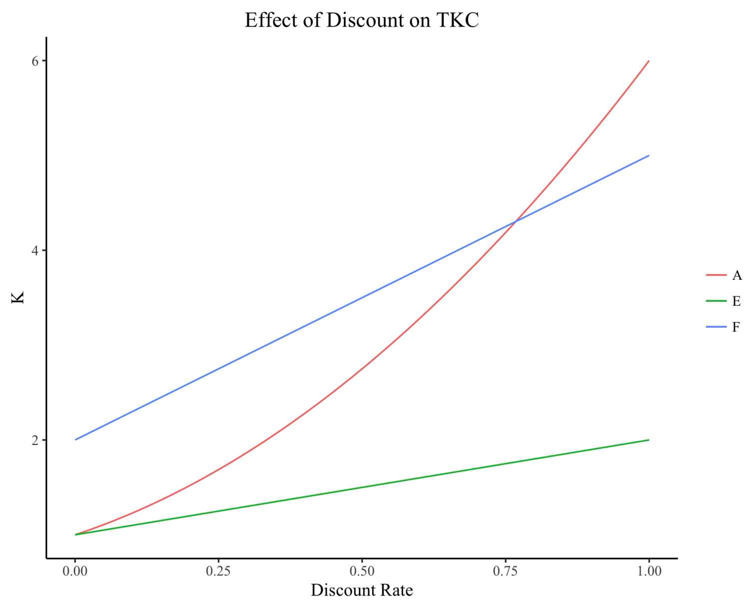


Figure 5: Total knowledge impact vs. discount rate for the highest contributors in the example network shown in Figure 3.

Though this approach better expresses total knowledge contribution intuitively, it necessarily falls short of a perfect description and is difficult to empirically verify since no ground-truth data exist. Because patent authors have varying thresholds for the strength of technological connection needed to include a citation and every author has imperfect information, we must assume that all patent authors are equally likely to cite a patent given the strength of its technological relation to their work. A correlated assumption is that all descendants are equally original; there must some constant ratio of the knowledge contribution of descendants stemming from its parents to the knowledge contribution stemming from original content in the descendant. It is quite possible that a patent's knowledge contribution is inflated because it happens to have a highly original, innovative descendant whose respective contribution is largely unrelated to its cited parent patent. Further, we must assume that each descendant's parent patents contribute equally, since it is impossible to discern the strength of contribution of each parent from citations alone.

Notwithstanding these drawbacks, the composite TKC index is a better theoretical descriptor of how much impact a given patent has on research and development after its publication. Empirically, the distribution of the TKC using both weighting methods is displayed in Figure 6 and described in Table 1. $K$ occupies a naturally logarithmic space, so the distributions are expressed on a logarithmic scale. Both distributions have extreme outliers - those highly prolific patents - and have roughly the same shape, though forward citations tend to magnify the TKC and reduce the number of outliers.

The means of all the test sets differ significantly at the 1% confidence level, with the exception of the x-ray technology group and the radio group (Table 2).

| Statistic | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|
| Forward Cites | 6.631 | 10.768 | 0 | 304 |
| Backward Cites | 6.631 | 11.947 | 0 | 404 |
| Family Size | 2.562 | 1.903 | $-5$ | 12 |
| Number of Claims | 1.998 | 0.571 | 0 | 4 |
| H Index | 2.537 | 3.296 | 0 | 54 |
| Knowledge (H Index) | 4.350 | 7.787 | 0.000 | 287.487 |
| Knowledge (Forward Cites) | 11.547 | 22.241 | 0.000 | 1,203.364 |

Table 1: Descriptive analytic statistics across all test data sets.

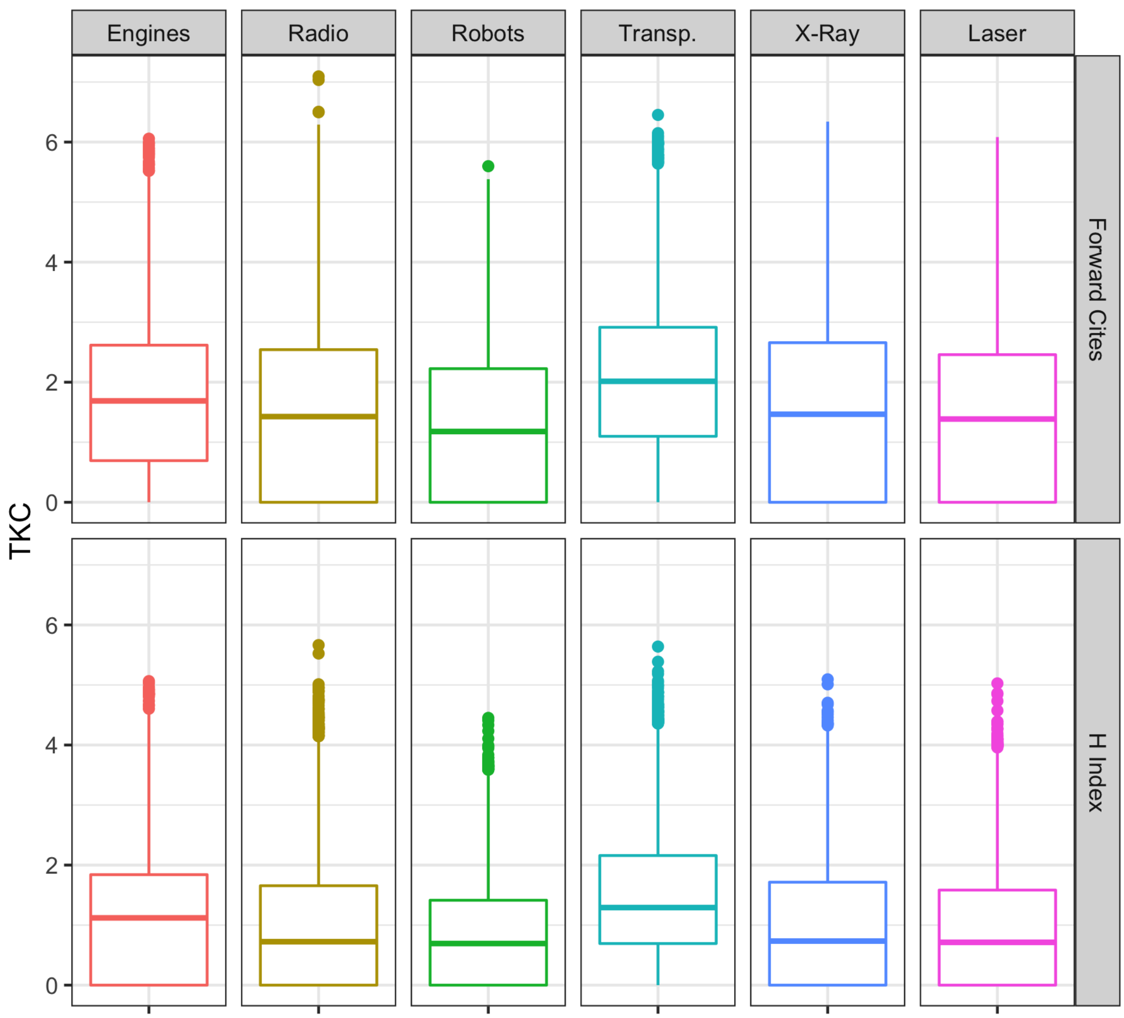


Figure 6: Distribution of the logarithm of total knowledge contribution calculated using two weighting methods for each test sector.

| | Engines | Radio | Robots | Transportation | X-Ray |
|---|---|---|---|---|---|
| Radio | 1.862E-08 | | | | |
| Robots | 6.478E-20 | 1.440E-31 | | | |
| Transportation | 1.011E-188 | 2.073E-68 | 2.414E-116 | | |
| X-Ray | 1.237E-11 | 1.486E-01 | 5.788E-35 | 8.699E-48 | |
| Coherent Light | 1.601E-04 | 7.052E-15 | 6.433E-10 | 1.019E-144 | 2.629E-18 |

Table 2: Pairwise t-tests with pooled standard deviation.

## 4 Predicting Knowledge Impact

Not only is the TKC useful for comparing stratifications of patents, it can also be applied to identify the observable predictors of impact of future innovation. We apply an ordinary least squares regression to evaluate the impact of available observable features of patents on their TKC indices and use a forecasting model to predict TKC for recently published patents. Finally, we attempt to measure time effects on cross-sectional patent time series with a pooled ordinary least squares model and quantify the impact of the AIA on patent knowledge propagation.

### 4.1 Predictors

We construct an ordinary least squares linear regression model:

$$TKC_i = \beta_0 + \beta_1 ln(claims_i) + \beta_2 ln(inventor\ patents_i) + \mathbf{B'}_i\mathbf{X}_i + e_i$$

where $TKC_i$ is the total knowledge contribution index for the i-th patent; $B$ is the exogenous coefficient matrix, including binary dummies representing all but one possible categorical value for the type of assignee and NBER category; $X$ is a vector of other observable controls.

We conduct a sequence of tests and interpret the statistics using conventional significance criterion. First, we find that assignee type and NBER dummies are jointly significant and include them in our specification. Second, we determine that patent processing time is significant, but find that its variance inflation factor is high and do not include it in our specification to minimize multicollinearity. (Processing time is correlated with the number of claims submitted as a pre-publication variable.) We also choose to include significant control variables, including the duration in days between January 4, 1976 and the date of publication and average assignee number of patents. The specification is also robust to dropping patents from any one assignee type. Next, we test the null hypothesis that an intercept-only model achieves has the same fit as our specification. This hypothesis is decidedly rejected ($p < .001$) by the F-statistic in 3. Finally, for each indicator, we test the null hypothesis that the coefficient is equal to zero. The results are reported in 3. We repeat the experiment for each technology subset, excluding the NBER category ID since there is little variation within technology groups.

These results provide some interesting insights into the components of total knowledge contribution. Increasing the number of claims by a one percent increases TKC by 0.34% (an increase in the number of claims from two to three increases TKC by nearly 17%, or 11.3% from three to four). Evidently, the number of claims a patent makes corresponds to its actual knowledge contribution by varying degrees in all technology fields, which aligns with prior research (Lanjouw and Schankerman, 2004). Furthermore, a percent increase the average number of patents held by the inventors corresponds to a 0.54% increase in TKC; in other words, an increase in the average inventor patent number by one from the mean 29 results in a 1.9% increase in TKC. More experienced inventors tend to produce patents that contribute more to future knowledge.

Remarkably, the type of patent assignee has significant relative effect on knowledge contribution. Compared to U.S. companies, the U.S. government produces patents with 122% higher TKC. U.S. individuals and foreign governments tend to produce patents with even higher TKC than either U.S. companies or the U.S. government. Other than U.S. companies, foreign companies produce the least impactful patents.

On the other hand, the NBER industry category of a patent has little to do with its TKC in general. Computer and communications patents tend to produce patents with 2% higher contribution than patents in the “Other" category, as do electronics patents.

## 5 Effect of Sectors & Policies

Evaluated over a patent's descendant tree at any given time, the TKC describes the extent of a patent's knowledge contribution at a stage in its development. We seek to establish a method for describing the normal progression of knowledge contribution through a patent's lifetime and identifying alterations to that evolution caused by particular policies or membership in a particular technology group.

For example, Figure 7 displays average knowledge contribution over time across technology sectors. The rate of contribution is fairly consistent across tech sectors. The Robots sector is particularly erratic; between 1984 and 1987, average knowledge contribution abruptly dips before reverting to the mean growth displayed by other sectors. Most likely, this aberration was caused by the advent of robotic surgery techniques in the mid-1980s. In general, a sharp decrease in average knowledge contribution indicates an influx of novel - new patents are added to the network without backward citations that would cause a commensurate increase in the TKC of older patents, thus reducing the mean TKC. We expect an steady increase in the mean TKC to follow as subsequent patents iterate on the novelties introduced during the negative growth period.

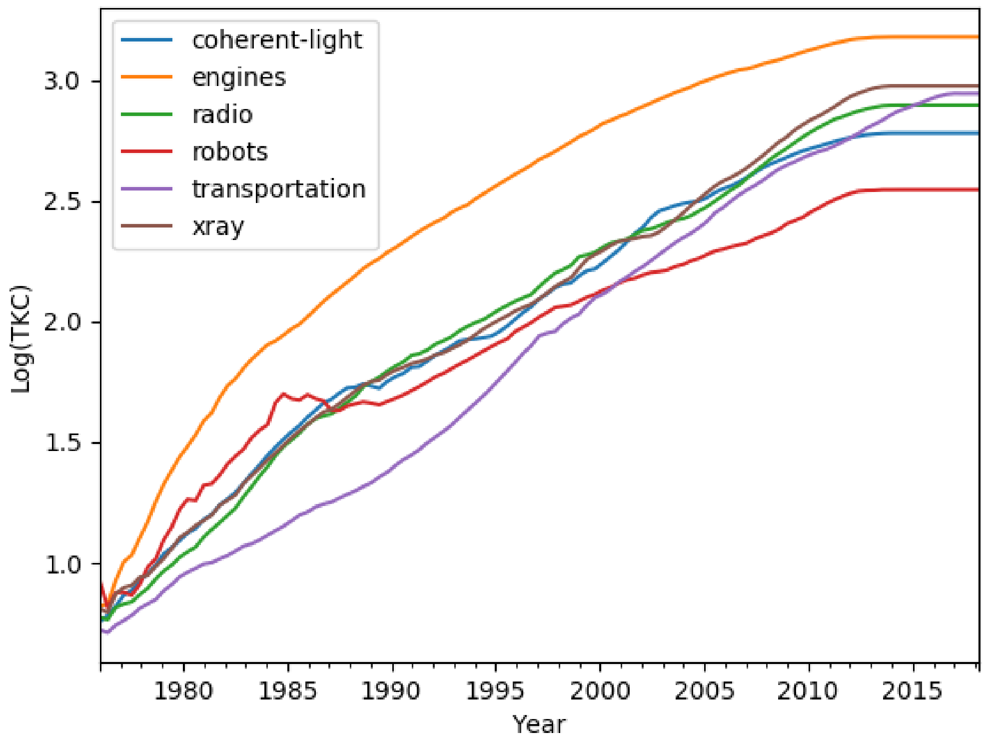


Figure 7: Mean knowledge contribution over time, stratified by technology sector. The flattening effect at the end of each curve is due to a reduction in patent addition rates towards the end of the time period examined (2016).

In this section, we present a case study based on the changes to the patent system wrought by the America Invents Act, which substituted a first-to-invent policy with a first-to-file policy. First, we estimate an autoregressive moving average (ARIMA) model on pre-AIA data to predict the "normal" or expected contribution trend before the policy was introduced in 2011. Because errors on a random walk model were autocorrelated, we select a differenced second-order model, or ARIMA(2,1,0), to best fit the data (lowest AIC score). Figure 8 depicts the forecast results. After 2014, the actual mean TKC significantly differs from the expected trend. A reduction in mean TKC may be due to an influx of new patents with fewer backward citations, or a reduction in patent production overall.

$$T\bar{K}C_t = \mu + T\bar{K}C_{t-1} + T\bar{K}C_{t-2} + \phi(T\bar{K}C_{t-1} - T\bar{K}C_{t-2})$$

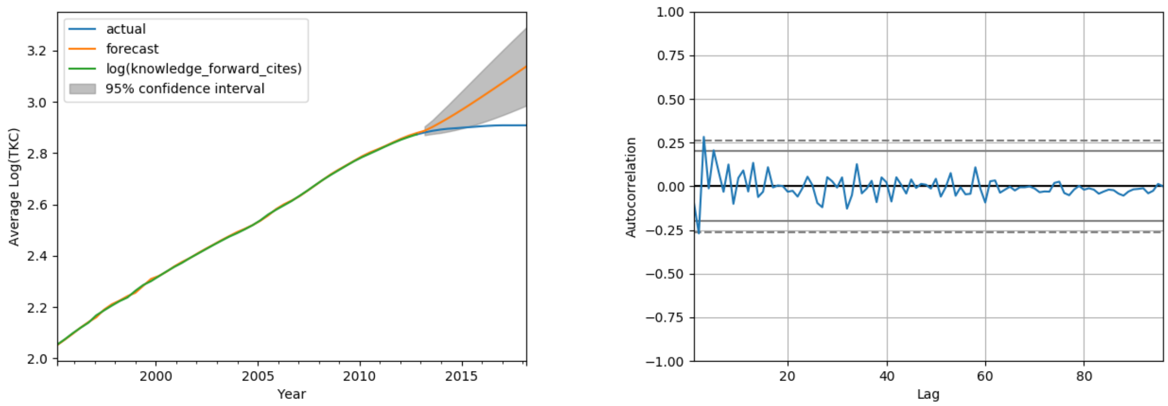


Figure 8: Average knowledge contribution ARIMA(2,1,0) forecast for all datasets, compared before and after the AIA effective date (left) and autocorrelation plot (left). A 95% confidence interval is shown for the forecast of TKC based on pre-AIA data. Only one point in the first 25 lags has significant autocorrelation at the 1% level, represented by a dotted line.

To further delineate the effects of time from the effects of exogenous factors and to account for latent differences between patents, we construct a pooled OLS regression on the time series cross-sectioned by patent. Time dummies are included to estimate the effect of each measured time period on total knowledge contribution, exogenous features and entity-specific variation held equal. We use clustered entity coefficients to account for variations between patents.

$$TKC_{it} = \beta_0 + \beta_1 TKC_{i,t-1} + \mathbf{B'}_i \mathbf{X}_i + \epsilon_{it}$$

Table 4 reports the exogenous parameter estimates in the pooled OLS model, which greatly reduced in comparison to the simple OLS model. In addition to the exogenous variables used before, we include patent age to account for the centered effect of time on TKC. Age accounts for a 16.7% increase in TKC per year. Sans patent age, the time dummy coefficients represent the latent effect of the time period on total knowledge contribution. This latent effect is a combination of the institutional factors that may cause new patents to contribute more or less to posterity. We theorize that these institutional factors are primarily composed of policy legislation, and that the introduction of major policy change such as the AIA will present itself in the effect measured by the time dummy variables in this model.

| | |
|---|---|
| **Constant** | 1.5374 (0.0564)*** |
| **Log(Number of Claims)** | 0.0574 (0.0060)*** |
| **Log(Avg. Inventor Patents)** | 0.1266 (0.0035)*** |
| **Patent Age (yrs)** | 0.1546 (0.0006)*** |
| **Patent Age Squared** | -0.0028 (0.0000)*** |
| **R-squared** | 0.3486 |
| **Prob (Robust F-statistic)** | 0.0000 |
| **No. Observations** | 3,731,942 |
| **Patents** | 63383 |
| **Avg. Obs.** | 58.879 |
| **Time periods** | 111 |

* $p < 0.05$, )* $p < 0.01$, *** $p < 0.001$.

Table 4: Pooled regression coefficients for TKC (forward cites). The patent processing time and dummy variables for the data source sector (engines, radio, etc.) and assignee type were also included in the regression specification but are not shown.

Figure 9 plots the time dummy coefficients. As expected, knowledge contribution increases with time. At its peak effect in 2010, TKC is nearly 50% higher than the intercept. But in late 2010 and early 2011, the effect of time begins to decrease sharply. In the pooled regression, the effects of patent productivity and patent age are eliminated, as well as the differences between individual patents. The decline can only be attributed to the economic and policy conditions governing the behavior of firms who submit patents.

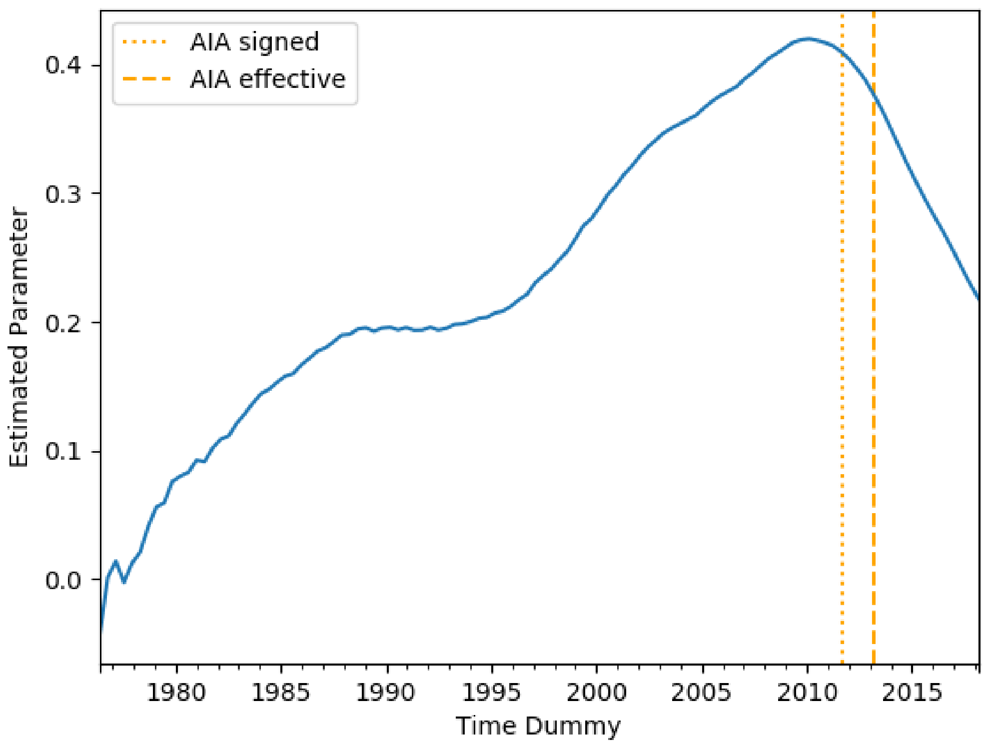


Figure 9: Parameter estimates for time dummies, smoothed.

The coincidence of the decline in TKC rates and the legislation of the AIA suggests that the change in policy had a negative effect on contribution rates. But why did the peak time effect occur before the legislation was signed or enforced? We hypothesize that because the AIA ostensibly increased intellectual property competition between firms, the anticipation of the proposed legislation caused firms to alter their patenting behavior as it was passed over the course of a year, filing for patents that were previously safe as trade secrets.

Why did the AIA cause contribution rates to decline? The first-to-file system encourages higher volume patenting, which would distribute knowledge across a larger number of patents and reduce the strength of contribution of individual patents. Some argue that the first-to-file system favors corporations with more patenting resources; we have shown that these types of entities produce patents with significantly lower contribution (Abrams and Wagner, 2013; Pierce, 2015).

## 6 Conclusions

Generally, the TKC index is a promising metric for evaluating knowledge flow networks of any kind. Improving on simple quality indices, the TKC more closely models the role of prior art in innovation. The TKC is especially useful for making statements about the factors which encourage contribution, especially government support and policy conditions. In the domain of intellectual property and innovation, we demonstrate the use of this framework for policy evaluation and explore some of its properties, including its distribution across sectors and its relationship to the major influences in the content, particularly the assignee. We find that while government patents tend to contribute more, recent legislation has had a negative effect on contribution rates, which are in decline.

Avenues of future research are numerous: investigate the external validity of TKC by applying it to other domains of knowledge flow; verify the internal validity of the TKC with a ground-truth variable or a latent factor model; investigate more sectors and identify the sector-wide determinants of knowledge contribution structures.

## 7 Acknowledgements

The author thanks Professor Rahul Simha for his mentorship and guidance. The USPTO PatentsView API was used to collect patent data. The Colonial One High Performance Computing cluster at the George Washington University was used for computation.

Software, algorithms, and analysis scripts used to collect these data and produce the figures in this work are open source and the command line API can be accessed at `rbsteed.com/datamaster`.

## References

Abrams, D. S. and R. P. Wagner (2013). Poisoning the Next Apple? The America Invents Act and Individual Inventors. *Stanford Law Review 65*(3), 517–563.

Chang, K.-C., W. Zhou, S. Zhang, and C.-C. Yuan (2015, 12). Threshold effects of the patent H-index in the relationship between patent citations and market value. *Journal of the Association for Information Science and Technology 66*(12), 2697–2703.

Daim, T. U., G. Rueda, H. Martin, and P. Gerdsri (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting & Social Change 73*, 981–1012.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001, 10). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Technical Report 8498, National Bureau of Economic Research.

Henrique, B. M., V. A. Sobreiro, and H. Kimura (2018, 5). Building direct citation networks. *Scientometrics 115*(2), 817–832.

Karanikić, P., I. Mladenović, S. Sokolov-Mladenović, and M. Alizamir (2017, 5). Prediction of economic growth by extreme learning approach based on science and technology transfer. *Quality & Quantity 51*(3), 1395–1401.

Lanjouw, J. O. and M. Schankerman (2004, 4). Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators. *The Economic Journal 114*(495), 441–465.

Leskovec, J., J. Kleinberg, and C. Faloutsos (2007, 3). Graph evolution. *ACM Transactions on Knowledge Discovery from Data 1*(1), 2–es.

Martinelli, A. and O. Nomaler (2014, 7). Measuring knowledge persistence: a genetic approach to patent citation networks. *Journal of Evolutionary Economics 24*(3), 623–652.

Miyagiwa, K. (2015, 6). The 2011 America Invents Act: Does it Undermine Innovation? *Journal of Economics & Management Strategy 24*(2), 211–227.

Oh, G., H.-Y. Kim, and A. Park (2017, 5). Analysis of technological innovation based on citation information. *Quality & Quantity 51*(3), 1065–1091.

Pierce, N. S. (2015). Inventorship, Double Patenting, and the America Invents Act. *Berkeley Technology Law Journal 30*(2), 1613–1686.

Pierce, S. (2012, 1). The Effect of the Leahy-Smith America Invents Act on Collaborative Research. *Journal of the Patent and Trademark Office Society 94*(2).

Strandburg, K. J., G. Csardi, J. Tobochnik, and P. Erdi (2009). Patent Citation Networks Revisited: Signs of a Twenty-First Century Change. Technical report, University of North Carolina.

Trajtenberg, M., R. Henderson, and A. Jaffe (1997). Economics of Innovation and New Technology University Versus Corporate Patents: A Window On The Basicness Of Invention. *Economics of Innovation and New Technology 5*(1), 19–50.

You, H., M. Li, K. W. Hipel, J. Jiang, B. Ge, and H. Duan (2017, 4). Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics 111*(1), 297–315.

| | | | Coherent Light | Engines | Radio | Robots | Transportation | X-Ray |
|---|---|---|---|---|---|---|---|---|
| **Log(Number of Claims)** | 0.3442 (0.007)*** | 0.3879 (0.003)*** | 0.9184 (0.012)*** | 1.0205 (0.006)*** | 1.0763 (0.012)*** | 0.9587 (0.018)*** | 0.9803 (0.023)*** | 1.1085 (0.013)*** |
| **Log(Avg. Inventor Patents)** | 0.5355 (0.003)*** | 0.5401 (0.003)*** | 0.4507 (0.010)*** | 0.3718 (0.005)*** | 0.3401 (0.010)*** | 0.2711 (0.015)*** | 0.4935 (0.006)*** | 0.3654 (0.011)*** |
| **Assignee: Foreign Co.** | 0.1619 (0.013)*** | 0.2022 (0.009)*** | 0.0957 (0.023)*** | 0.3052 (0.012)*** | 0.2098 (0.024)*** | 0.2806 (0.038)*** | 0.3887 (0.017)*** | 0.1403 (0.026)*** |
| **Assignee: U.S. Ind.** | 0.8658 (0.051)*** | 0.8890 (0.051)*** | 0.6161 (0.360) | 0.5628 (0.081)*** | 0.6909 (0.187)*** | 0.7913 (0.424) | 0.1493 (0.081) | 0.4896 (0.148)** |
| **Assignee: Foreign Ind.** | 0.0940 (0.070) | 0.1266 (0.071) | -0.3484 (0.257) | 0.3607 (0.088)*** | -0.1297 (0.191) | 0.2635 (0.264) | -0.5221 (0.157)*** | -0.0010 (0.208) |
| **Assignee: U.S. Govt.** | 0.7993 (0.031) | 0.7568 (0.030)*** | 0.5339 (0.51)*** | 0.4219 (0.163)** | 0.6432 (0.040)*** | 0.5605 (0.141)*** | -1.1716 (0.269)*** | 0.6926 (0.116)*** |
| **Assignee: Foreign Govt.** | 0.8674 (0.076)*** | 0.8275 (0.077)*** | 0.5635 (0.122)*** | 0.9560 (0.145)*** | 0.5816 (0.141)*** | 1.0516 (0.184)*** | -3.1230 (0.051)** | 0.7443 (0.166)*** |
| **NBER: Chemical** | -0.0532 (0.043) | | | | | | | |
| **NBER: Computer & Comm.** | 0.0295 (0.020) | | | | | | | |
| **NBER: Drugs & Medical** | 0.3032 (0.063)*** | | | | | | | |
| **NBER: Electronics** | 0.0201 (0.020) | | | | | | | |
| **NBER: Mechanical** | 0.2396 (0.018)*** | | | | | | | |
| **Adj. R-squared:** | 0.724 | 0.723 | 0.794 | 0.802 | 0.754 | 0.774 | 0.797 | 0.768 |
| **Prob (F-statistic)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **No. Observations:** | 133760 | | 14741 | 48580 | 17552 | 4458 | 34710 | 13719 |

* $p < 0.05$, )* $p < 0.01$, *** $p < 0.001$. Joint p-values calculated with an F-test of $H_0 : (\forall\beta_i)(\beta_i = 0)$.

Table 3: TKC (forward cites) regression coefficients using all test data sets. Robust standard errors are reported in parentheses. (*) indicates statistical significance at the 1% level ($p < .001$). Note that dummy variables for all assignee type and NBER category id were included in the regression specification.