Searching for Innovation:

Understanding Patent Impact with Google's PageRank Algorithm

David S. Abrams* Elliott Ash†

June 24, 2026

Abstract

General-purpose proxies for the value of innovation are central to a wide range of research questions, yet the dominant proxy—forward citation counts—treats every citation as equivalent. We propose a family of patent-impact metrics built on PageRank, the algorithm behind Google search, which weights each citation a patent receives by the importance of the citing patent and thus captures indirect influence through citation chains. We compute PageRank and its age-adjusted and age-and-technology-adjusted variants for the universe of roughly 12 million US granted patents, and make them freely accessible. The metrics are strongly correlated with forward citations and with stock-market value, but concentrate the inventions that external sources regard as foundational more tightly in their upper tail. The measure also recovers underappreciated patents—inventions with few citations but high PageRank, because the few patents that cite them are themselves landmarks. Holding citations fixed, such patents are more likely to be text-based breakthroughs. Finally, we revisit the effect of patent term on the value of innovation.

*University of Pennsylvania Law School. dabrams@upenn.edu.
†ETH Zürich Center for Law and Economics. ashe@ethz.ch.

1 Introduction

Which inventions matter, and how much? A usable, general-purpose proxy for the value of an individual invention underlies a wide range of questions in economics: the returns to R&D, the sources and direction of technical change, knowledge spillovers, the design of intellectual-property policy, and the measurement of firm and regional innovativeness. Patents provide a rich, publicly available record of inventive activity that researchers have exploited since at least Griliches (1990), but patents vary enormously in significance—fewer than ten percent are worth the fees required to renew them, while a handful are worth hundreds of millions of dollars (Pakes, 1986; Hall et al., 2005; Abrams et al., 2013). A simple patent count is therefore a very noisy proxy for inventive output.

The standard refinement is to weight patents by the forward citations they receive. Citation-weighted counts became the workhorse measure of patent quality after Trajtenberg (1990), and have been validated against firm market value (Hall et al., 2005), grant-day stock-market reactions (Kogan et al., 2017), and inventor surveys (Jaffe et al., 1993). Yet a citation *count* has a basic limitation: it weights every citation equally. A citation from a foundational, much-built-upon patent counts exactly the same as one from a patent that itself went nowhere. This single feature drives three familiar problems—counts are inflated in high-volume, densely citing recent fields; they are padded by examiner-added and strategic citations that need not reflect genuine knowledge flows (Abrams et al., 2013; Jaffe and de Rassenfosse, 2017); and they reward the sheer quantity of attention over position in the web of innovation. Stock-market-based value (Kogan et al., 2017) and text-based breakthrough measures (Kelly et al., 2021) are valuable complements, but neither scores the whole corpus: the former requires a publicly traded assignee whose stock reaction can be isolated, the latter comparable full text and a chosen similarity window.

A natural alternative comes from the PageRank algorithm. PageRank is itself a patented invention—US Patent 6,285,999, "Method for node ranking in a linked database," assigned to Stanford University—and it is one of the breakthroughs that curated histories of technology single out (Kelly et al., 2021). Its central idea is the one a citation count lacks: rank a node not by how many links it receives but by how important the linking nodes are, computed recursively.

We bring this idea to patents. We propose a family of patent-impact metrics built on PageRank (Brin and Page, 1998) and compute them for the universe of US granted patents. PageRank defines the importance of a patent recursively: a patent is important if it is cited by other important patents. Unlike a raw count, it weights each received citation by the importance of the citing patent and discounts citations from patents that themselves cite

indiscriminately, so it captures indirect influence through citation chains—a patent cited by influential follow-on work occupies a structurally different position than one cited the same number of times from the periphery. The approach has identified important scientific papers (Mariani et al., 2019; Xu et al., 2022) and biotechnology patents (Reinstaller and Reschenhofer, 2017), but has not been implemented at the scale of the full US corpus or released as a ready-to-use measure. We compute PageRank on the US-grant-to-US-grant citation network (about 12 million patents) and release three variants—raw PageRank, an age-adjusted variant that removes the mechanical advantage of older cohorts, and an age-and-technology-adjusted variant that additionally normalizes within field—derived entirely from public data with open-source code.[1]

These metrics are strongly correlated with forward citations and with stock-market value, yet appear better at identifying the top inventions. Benchmarked against external ground truth—KPSS private value (Kogan et al., 2017), KPST text-based breakthroughs (Kelly et al., 2021), and curated lists of landmark inventions—PageRank concentrates the patents that independent sources regard as foundational more tightly in its upper tail than raw citation counts do. The advantage is sharpest for scientific and technological importance; the association with private market value, by contrast, is weaker and sensitive to specification, underscoring that network centrality and private appropriability are distinct dimensions of a patent's value.

A further payoff of a network measure is the class of patents it recovers from citation obscurity. Some inventions receive few citations yet sit upstream of major advances, because their handful of citers are themselves landmarks. A citation count cannot identify them, whereas PageRank can. We show that, among US patents with at most the median number of forward citations, those PageRank nonetheless ranks highly are about 1.8 times as likely to be independent text-based breakthroughs as citation-matched peers. These under-appreciated patents are the kind of foundations an impact measure should surface. Finally, because the measure is as easy to use as a citation count, it is a practical drop-in for the many applied designs that currently rely on citations; we illustrate this by revisiting the effect of patent term on the value of innovation (Abrams, 2009).

Section 2 reviews the related literature. Section 3 describes the data. Section 4 presents the PageRank measurement framework. Section 5 reports summary statistics and validation. Section 6 develops the under-appreciated-patents result, and Section 7 applies the measure to patent term. Section 8 concludes.

[1]The full prior-art graph that also includes US applications and foreign cited publications is analyzed in the supplement; the main specification is the US-grant-only graph.

2 Related Literature

Patent Citations as Indicators of Innovative Value. The systematic use of patent data in economic research goes back at least to Griliches (1990), who surveys the literature on patent statistics as economic indicators and emphasizes both the promise and the limitations of patent counts. The critical insight that patent *citations* could improve on patent counts as a value indicator was developed by Trajtenberg (1990), who shows that in the computed tomography industry, citation-weighted patent measures track independently estimated social values of innovations far better than unweighted counts. Jaffe and Trajtenberg (2002) provide the conceptual framework and comprehensive dataset that established patent citation analysis as a standard tool in economics.

The literature has confirmed the citation–value relationship using a variety of value proxies. Harhoff et al. (1999) use survey data on German patent holders to show a strong positive correlation between forward citation counts and inventor-reported patent value. Hall et al. (2005) construct a landmark dataset of US patents from 1963–1999 and demonstrate that citation-weighted patent stocks predict firm market value (Tobin's q) significantly better than raw patent counts, with each additional citation per patent associated with a 3% increase in firm market value. Kogan et al. (2017) develop a complementary stock-market-based measure of patent importance using daily stock returns around patent grant announcements and find it positively correlated with forward citations while containing additional information about subsequent firm growth.

University patents and knowledge spillovers have received particular attention in this literature. Henderson et al. (1998) show that university patents were more highly cited than corporate patents in the pre-1982 period, with implications for technology transfer policy. Jaffe et al. (1993) use the geographic co-location of cited and citing patents as evidence of geographically localized knowledge spillovers, a finding that has shaped decades of work on the spatial economics of innovation.

Heterogeneity and Validation of Patent Quality Measures. Patent value distributions are highly skewed. Silverberg and Verspagen (2007) apply extreme value statistics to both citation and survey-based value measures, estimating Pareto tail exponents in the range 1–2, confirming the power-law-like behavior of the upper tail. Lanjouw and Schankerman (2004) construct composite quality indicators using multiple patent attributes—forward citations, backward citations, family size, claims, and renewals—and find that forward citations remain among the most informative individual predictors of patent quality.

Recent work has complicated the simple monotone relationship between citations and

value. Abrams et al. (2013) use proprietary licensing revenue data for tens of thousands of patents held by non-practicing entities and document an inverted-U relationship: while citations predict value positively for lower-value patents, the most valuable patents in the upper tail receive *fewer* citations than moderately valuable ones. They attribute this to strategic patenting behavior—high-value patents create incentives to erect barriers through "fencing" patents that block downstream innovation without generating knowledge spillovers. Higham et al. (2021) propose a systematic framework for patent quality measurement, arguing that different indicators capture distinct dimensions of quality and advocating for multi-indicator approaches in empirical research.

Jaffe and de Rassenfosse (2017) provide a comprehensive review of best practices for using patent citation data in social science research, covering critical methodological issues including the distinction between examiner-added and applicant-added citations, time-varying citation norms across fields, and right-truncation of citation counts for recent patents. We follow their recommendations in our data construction and analysis.

Network Centrality for Citation Impact Measurement. The application of network centrality methods to citation graphs has a long history in the scientometrics literature. Verspagen (2007) uses citation network analysis to map the trajectory of fuel cell technology, demonstrating that the network structure of citations reveals knowledge-flow patterns that raw counts alone cannot capture.

The PageRank algorithm, introduced by Brin and Page (1998) as a method for ranking web pages by their recursive hyperlink importance, is particularly well suited to citation networks. Like web pages, patents can be thought of as nodes in a directed network where the importance of each node depends on the importance of the nodes that point to it. The algorithm's main advantage over raw in-degree (citation count) is that it assigns higher weight to citations from central, highly-cited patents.

In the economics and management of technology literature, Reinstaller and Reschenhofer (2017) apply PageRank to biotechnology patent citations and find that, with appropriate calibration, it produces impact rankings that improve on raw citation counts as predictors of expert assessments of technological importance. Mariani et al. (2016) develop time-balanced network centrality for identifying milestone scientific papers. Mariani et al. (2019) adapt that logic to the US patent citation network (1926–2010), showing that time-rescaled PageRank improves the early identification of expert-selected important patents relative to raw citation counts. Xu et al. (2022) provide a comprehensive unbiased comparison of seventeen network-based ranking metrics on citation datasets from physics, high-energy physics, and the US patent office, concluding that PageRank (and the related LeaderRank) consistently

outperforms citation count when properly corrected for age bias.

Text-Based Measures of Patent Importance. A parallel literature measures patent importance from document text rather than citations. Younge and Kuhn (2016) construct pairwise patent-to-patent textual similarity at scale; Ashtor (2019) uses cohort textual similarity as an ex-ante alternative to forward citations; and Kelly et al. (2021) identify breakthrough patents as those textually novel relative to prior art yet similar to subsequent work, validating against historical landmarks and market value much as we do. These text-based measures and our network-based measure are complementary routes to the same goal—improving on the raw citation count—and we use the Kelly et al. (2021) breakthrough indicator as one of our external benchmarks.

We build on this literature primarily in scale and accessibility. We compute PageRank for the universe of US granted patents on the US-grant-to-US-grant citation graph (about 12 million patents and 160 million edges), release age- and field-adjusted variants, and validate them against curated landmarks, market value, and text-based breakthroughs, with full reproducibility via public data and open-source code. Crucially, we show what the network measure recovers that a citation count cannot: under-appreciated patents whose few citers are themselves important (Section 6).

3 Data

Google Patents Public Data. Our primary data come from the Google Patents Public Data corpus, hosted as a publicly available table on Google BigQuery. The patents table contains bibliographic information on patents and patent applications from more than 100 patent offices worldwide, maintained by Google in collaboration with IFI CLAIMS Patent Services. Each record includes the publication number, filing date, grant date, inventor names, assignee information, technology classifications (Cooperative Patent Classification [CPC], International Patent Classification [IPC], and US Patent Classification [USPC]), and a structured array field listing the publication numbers of all prior patents cited by the document.

The corpus provides near-complete coverage of all grants issued by the United States Patent and Trademark Office (USPTO) from 1790 to the present, with machine-readable citation data reliably available from the mid-1970s onward. Unlike the historical NBER Patent Citations Data File (Hall et al., 2001), which covers patents granted through 1999, the Google Patents corpus is updated regularly and covers the full contemporary patent record. All data used in this study are publicly available at no cost via the Google BigQuery

free tier (up to 1 TB of queries per month).

Historical onset and time span of the citation data. Although the corpus extends back to 1790, US patents did not record citations to prior art for most of that history. References to earlier patents first appeared on the printed front page of US patents in 1947, when examiners began listing the prior art considered during examination; coverage broadened over the following decades, and citation data become machine-readable and reliably complete only from the mid-1960s—the starting point of the NBER Patent Citations Data File (Hall et al., 2001), which compiles citations from patents granted 1963–1999—while the USPTO began separately flagging examiner- versus applicant-supplied citations only in 2001 (Jaffe and de Rassenfosse, 2017). Backward citations are therefore essentially absent for patents granted before the late 1940s and sparse until the 1970s, even though those early patents still accrue forward citations from later filings. Our analysis draws on the full US grant history in the corpus, from the first grants in 1790 through the June 2026 extraction. We focus on the US-grant-to-US-grant citation graph and treat the period from the mid-1970s onward—when citation practice is well established—as the effective analysis window. we retain the older grants because the recursive PageRank computation relies on the early cited patents that anchor many citation chains. Further caveats on citation sources, right-censoring, and strategic citation are detailed in Appendix A.

The citation graph. After cleaning, the US-grant-to-US-grant network—a single, homogeneous population in which one US grant cites another—contains 12.0 million nodes and 160 million directed edges. Table 1 reports these counts together with the PageRank parameters used throughout. The extraction query, edge-cleaning steps, and the larger full prior-art graph—which additionally includes US published applications and foreign cited publications (32.6 million nodes, 353 million edges)—are described in Appendix A.

Patent metadata. For the empirical analysis we attach two metadata fields from the same BigQuery table to each patent. The first is the *filing year*, used for the age adjustment. We also use the Cooperative Patent Classification (CPC) code as a measure of technological area. As is standard, we use the eight substantive CPC sections (A–H) and set aside the supplementary "Y" tagging scheme (sparse cross-sectional codes such as climate-change `Y02` and the former-USPC `Y10` cross-reference collections, under 0.01% of patents). Grouping patents by the first four characters of their primary CPC code (e.g., `H04L`) then gives 697 distinct CPC groups. The sections underlie the age-and-technology adjustment, while these finer groups are used for field comparisons. Construction details and the CPC sections are

Table 1: Citation Statistics and PageRank Parameters

Statistic	Value
US granted patents (nodes)	12,017,632
US→US citation edges	159,988,382
Damping factor d	0.85
Convergence tolerance (τ)	10^{-6} (absolute L1)
Iterations to convergence	29
Final abs-L1 delta	6.45×10^{-7}

Notes: Main specification: the citation graph is restricted to US granted patents on both the citing and cited side (US grant cites US grant). Full prior-art-graph statistics (including US applications and foreign cited publications) are reported in the supplement.

given in Appendix A.

4 Measuring Patent Impact with PageRank

This section lays out our approach for measuring patent impact using the PageRank algorithm.

Graph Formulation. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph where each node $v \in \mathcal{V}$ represents a unique patent publication number and each directed edge $(u, v) \in \mathcal{E}$ represents a citation from patent u to patent v. We follow the standard convention that edges point *from* the newer citing patent *to* the older cited patent, so that the graph is mostly (though not strictly) acyclic in time. A patent's "impact" is thus captured by the structure of edges pointing toward it: the set of patents that cite it.

The unit of analysis is the granted-patent publication number. The main specification restricts the graph to US granted patents on both sides, so published applications do not enter as separate nodes; canonical application-level collapsing of grants remains a natural extension. The full prior-art graph (which does include applications and foreign cited publications) is analyzed in the supplement.

The in-degree of node v, $k_{\text{in}}(v) = |\{u : (u, v) \in \mathcal{E}\}|$, is the raw forward citation count—the most widely used patent impact measure in the existing literature. The out-degree $k_{\text{out}}(v)$ counts the number of backward citations made by patent v.

The PageRank Algorithm. The PageRank algorithm (Brin and Page, 1998) defines a node importance score $r(v)$ as the stationary distribution of a random walk on the graph $\mathcal{G}$. A random walker positioned at node u at time t moves to time $t+1$ according to the following rule:

- With probability d (the *damping factor*), the walker follows a uniformly random outgoing edge from u, moving to one of u's out-neighbors.
- With probability $1-d$, the walker "teleports" to a node chosen uniformly at random from all $N = |\mathcal{V}|$ nodes in the graph.

The teleportation step is the key regularization that ensures the random walk has a unique stationary distribution even on disconnected or acyclic graphs. It can be interpreted as the probability that a reader of patents abandons the citation trail and opens a random patent from the full corpus. The damping factor d controls the balance between following citations and teleporting; we use the standard web-search value of $d = 0.85$ (Brin and Page, 1998), while reporting robustness to the lower $d = 0.5$ value used by Mariani et al. (2019).

The stationary distribution satisfies the fixed-point equation:

$$r(v) = \frac{1-d}{N} + d \sum_{u:\,(u,v)\in\mathcal{E}} \frac{r(u)}{k_{\text{out}}(u)} \tag{1}$$

for all $v \in \mathcal{V}$. The scores sum to one: $\sum_v r(v) = 1$. A high value of $r(v)$ indicates that patent v is frequently visited by the random walk—equivalently, that it is cited by many other important patents. We solve this fixed-point equation by power iteration. The details on the convergence criterion and the runtime of our implementation are detailed in Appendix A.

PageRank versus Raw Citation Count. For a node v, the raw in-degree $k_{\text{in}}(v)$ counts how many patents cite v, weighting all citations equally. PageRank $r(v)$ instead weights each incoming citation by the importance of the citing patent, which itself depends on the importance of its own citing patents, and so on recursively.

The key differences between the two measures can be seen from Eq. (1):

1. **Citation source quality.** A citation from a patent with high $r(u)$ contributes more to $r(v)$ than one from a patent with low $r(u)$. PageRank thus assigns higher weight to being cited by influential follow-on patents than to being cited by peripheral ones.
2. **Citation source quantity.** A citation from a patent with high out-degree $k_{\text{out}}(u)$ contributes *less* than one from a more focused patent that cites fewer works. This

downweights the contribution of patents that cite very large numbers of prior works, which may indicate less selective reference behavior.

3. **Network position, not just neighborhood.** Because the recursion runs to convergence, $r(v)$ captures the full network structure—the "prestige" of a patent's entire upstream citation tree—not just its immediate in-degree.

In the limit as $d \to 0$, all patents receive the same score $1/N$ (all teleportation, no citation following). As $d \to 1$, the algorithm approaches standard eigenvector centrality with no teleportation. At $d = 0.85$, the score is a weighted blend: 15% "baseline" from teleportation and 85% from following citations from important patents.

Percentile Divergence. To measure the degree to which individual patents are differently valued by PageRank versus raw citation count, we construct a *percentile divergence* measure:

$$\Delta_v = P_{PR}(v) - P_{\text{cit}}(v) \tag{2}$$

where $P_{PR}(v) \in [0, 100]$ is the percentile rank of patent v in the PageRank distribution (higher is more impactful) and $P_{\text{cit}}(v)$ is the corresponding percentile rank in the raw citation-count distribution. A positive value $\Delta_v > 0$ indicates that PageRank rates the patent as more important than raw citations would suggest; a negative value indicates the reverse.

This measure is useful for identifying which types of patents gain or lose impact when network position, rather than raw citation count, is taken into account. A patent at the center of a highly cited technology cluster may have $\Delta_v \gg 0$ even with moderate raw citations; conversely, a patent with many citations from scattered, low-influence work may have $\Delta_v \ll 0$. We exploit this measure in the empirical analysis (Section 5) to understand the economic content of the divergence between the two rankings.

Age and technology normalization. Raw PageRank is mechanically biased toward older patents: citations accumulate over time and the corpus has grown, so older cohorts have had longer to acquire the incoming paths the algorithm rewards. Following Mariani et al. (2019), we remove this bias by normalizing each patent's score within its filing-year cohort. Our headline age-adjusted variant is the cohort-mean rescaling

$$\tilde{r}_{\text{mean}}(v) = \frac{r(v)}{\bar{r}(t_v)}, \qquad \bar{r}(t) = \mathbb{E}[r(v) \mid t_v = t]\,, \tag{3}$$

where t_v is the filing year of patent v; a value of 2 means twice the impact of the typical patent filed the same year, and $\mathbb{E}[\tilde{r}_{\text{mean}} \mid t_v = t] = 1$ for every t once age bias is removed.

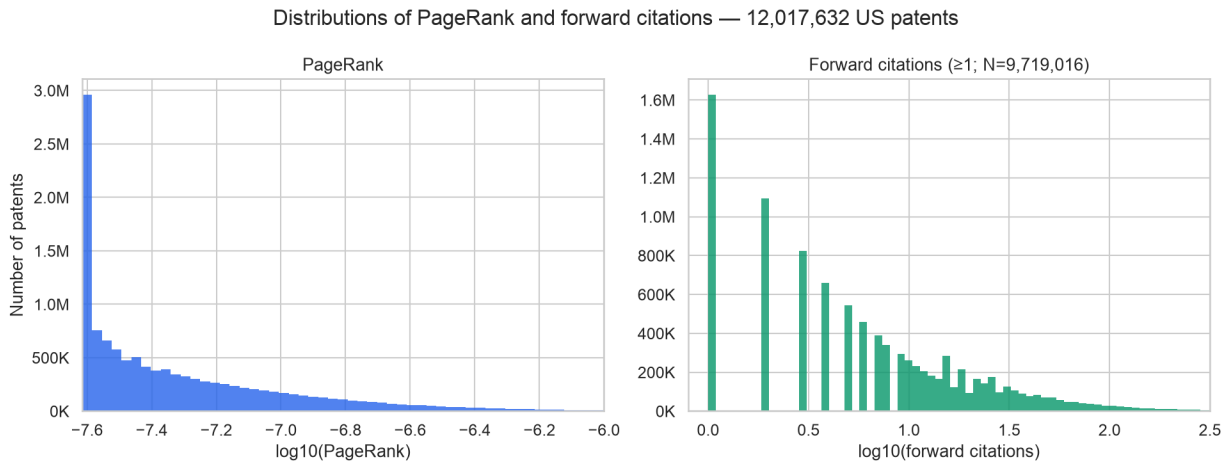


Figure 1: Distributions of PageRank and Forward Citations

Notes: Left: $\log_{10}$(PageRank) across all 12.0 million US grants (x-axis clipped to the populated range). Right: $\log_{10}$(forward citations) among grants with at least one forward citation. Both distributions are heavy-tailed.

We also release two robustness variants: a within-cohort log z-score $\tilde{r}_z(v) = (\ln r(v) - \mu_{\log}(t_v))/\sigma_{\log}(t_v)$, and the distribution-free cohort percentile rank $\tilde{r}_{\text{pct}}(v) = \Pr[r(v') \le r(v) \mid t_{v'} = t_v] \times 100$.

Cross-field differences in citing intensity—some technologies cite far more densely than others—bias cross-section comparisons in the same way age does. The *age-and-technology-adjusted* variant therefore normalizes within filing-year × CPC-section cells,

$$\tilde{r}^{\text{tech}}_{\text{mean}}(v) = \frac{r(v)}{\bar{r}(t_v, s_v)}, \qquad \bar{r}(t, s) = \mathbb{E}[r(v) \mid t_v = t,\ s_v = s]\,, \tag{4}$$

where s_v is the CPC section (A–H) of patent v and cells with fewer than twenty patents are left unscaled. This is our preferred measure for comparing patents across both vintage and technology field. All three normalized scores are released alongside raw PageRank.

5 Summary Statistics and Validation

5.1 Summary Statistics on PageRanked Patents

The distribution of patent importance. Figure 1 shows the distributions of PageRank and forward citations across the 12.0 million US granted patents in the main graph. Both are heavily right-skewed on a $\log_{10}$ scale: the modal patent sits near the teleportation floor

$1/N$, while a small number of foundational patents score several orders of magnitude higher. The PageRank mass is extremely concentrated—a Pareto-like upper tail dominated by a few computing, semiconductor, and biotechnology patents (the rank-size and Lorenz curves are in Appendix Figure A.1).

Age adjustment. The age-adjusted and age-and-technology-adjusted variants defined in Section 4 (Eqs. 3–4) remove the mechanical advantage of older, and more densely citing, cohorts. Appendix Figure A.3 shows the correction works: mean raw $\log_{10}$(PageRank) declines steeply with filing year, while the cohort-rescaled mean is flat at one and the cohort percentile is flat at fifty by construction. The age-and-technology variant additionally flattens cross-field differences (Appendix Figure A.8); this adjustment is what allows recent landmark inventions to surface against older cohorts.

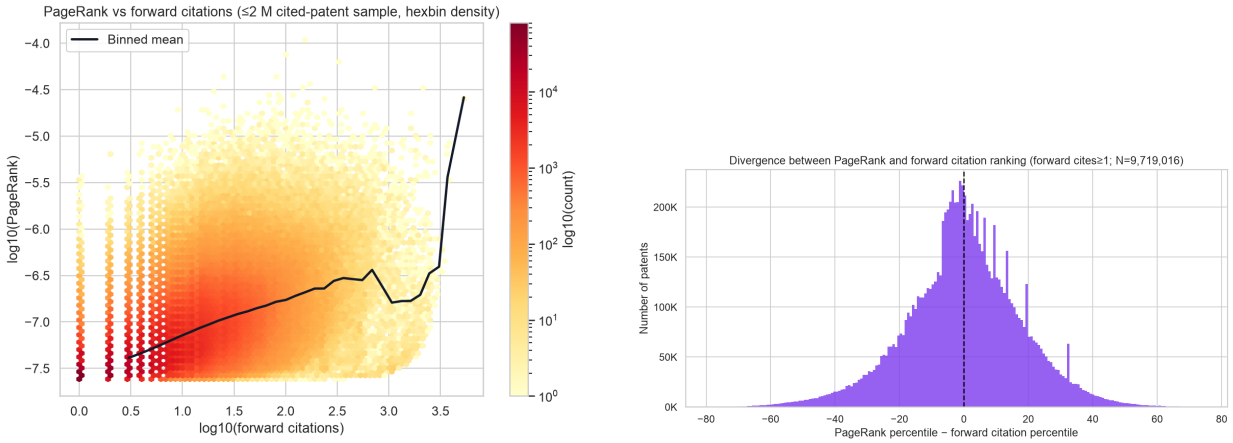
5.2 Validation: Does PageRank Identify the Important Patents?

We validate the measure with three checks, in the spirit of Kelly et al. (2021): against forward citations (the incumbent measure), against curated landmark inventions, and against external value measures (KPSS market value and KPST text-based breakthroughs). A recurring theme is that PageRank is most informative about *scientific and technological* importance, a dimension distinct from private market value.

Forward citations. PageRank and the raw forward-citation count are strongly positively related but far from identical. Figure 2a plots $\log_{10}$(PageRank) against $\log_{10}$(forward citations) for a two-million-patent sample: the binned-mean trend rises monotonically, but with wide scatter at every citation level, so network position carries information beyond the raw count.

The reranking is quantified by the percentile divergence $\Delta_v = P_{\mathrm{PR}}(v) - P_{\mathrm{cit}}(v)$ (Eq. 2). Figure 2b shows its distribution is centered near zero but wide and mildly asymmetric: many patents move substantially between the two rankings. Patents ranked far higher by PageRank tend to be cited by a few highly influential patents; those ranked higher by citations accumulate many citations from peripheral work. (The percentile scatter and the extreme-case table are in Appendix Figure A.2 and Table A.2; the largest within-decade divergences— overall and among the top-1% of patents by PageRank—are in Appendix Tables A.3 and A.4.)

The divergence is most pronounced at the top of the distribution. Table 2 lists the top-20 US patents by forward citations and by age-adjusted PageRank side by side: the two lists do not overlap. Forward citations surface recent, heavily cited patents, while age-adjusted



(a) PageRank vs. forward citations

(b) Divergence between the two rankings

Figure 2: PageRank versus Forward Citations

Notes: (a) Hexbin density of $\log_{10}$(PageRank) against $\log_{10}$(forward citations) for a 2-million-patent sample, with a binned-mean trend (black); positive but with wide scatter at every level. (b) Distribution of $\Delta_v = P_{\mathrm{PR}} - P_{\mathrm{cit}}$ over grants with at least one forward citation; positive values are patents ranked higher by PageRank than by raw forward citations (dashed line at zero).

Table 2: Top-20 US Patents: Forward Citations vs. Age-Adjusted PageRank

	By forward citations				By age-adjusted PageRank $\tilde{r}_{\mathrm{mean}}$			
Rank	Patent	Grant	c	Product	Patent	Grant	$\tilde{r}_{\mathrm{mean}}$	Product
1	PCR DNA amplification	1987	5888		Early transistorized computer	1962	647.9	MOBIDIC
2	Nucleic acid amplification	1987	5321		Computer data processing system	1965	540.9	
3	Petunia dwarfism gene	1996	5006		Multiplying and dividing machine	1952	451.9	
4	Monoclonal antibody production	1989	4649		Electric current control	1930	431.3	
5	Semiconductor device fabrication	2010	4199		Recombinant DNA	1980	392.0	
6	Electrosurgical sealing device	1995	3968	LigaSure	Information handling device	1962	363.7	
7	Single-crystal thin-film material	2006	3939		Floating-gate MOS transistor	1970	345.1	
8	Liquid crystal display fabrication	1998	3927	flat-panel display	Computer control circuits	1957	329.8	
9	Semiconductor device fabrication	2010	3924		Source code version control	1985	326.7	
10	Semiconductor device	2007	3910		Automatic code compilation	1982	325.2	
11	Robotic surgical tool	2004	3886	da Vinci surgical robot	Microbes with degradative plasmids	1974	312.1	
12	Oxide thin film	2001	3883		Electric current control	1933	306.5	
13	Semiconductor transistor	2004	3883		Picture transmission	1930	295.6	
14	Thin-film transistor display	2003	3865		Maser	1960	293.9	maser
15	Oxide semiconductor device	2007	3860		Atomic clock	1958	279.5	
16	Semiconductor transistor	2006	3849		Automatic message switching	1954	277.1	
17	Zinc-oxide transistor	2008	3838		Web browser linking display	1996	274.6	web browser
18	Single-crystal thin-film material	2008	3835		Liquid-crystal display element	1967	261.9	
19	Field-effect transistor	2008	3834		First transistor	1950	248.3	
20	Indium oxide transistor	2007	3829		Information handling system	1962	242.7	

Notes: Left block: top-20 US patents by forward citations (c). Right block: top-20 by age-adjusted PageRank ($\tilde{r}_{\mathrm{mean}}$, Mariani cohort-mean rescaling). Patents appearing in both lists would be bold-faced (none do). A recognized product/technique name is shown where known, otherwise the patent title.

PageRank surfaces a different set of network-central inventions—evidence that the measures identify different "top" patents.

Market and scientific value. We benchmark PageRank against two external value measures: the KPSS stock-market patent value (Kogan et al., 2017) and the KPST text-based breakthrough measure (Kelly et al., 2021). Table 3 reports both sets of regressions: Panels A–B take KPSS market value as the dependent variable, and Panels C–D take the KPST text-importance score.

For KPSS market value, the PageRank variants are positively associated with value under year and CPC fixed effects (Panel A); dropping those fixed effects collapses—and even flips—the raw-PageRank association, while the age-adjusted and R_p variants stay positive. In the citation horse race (Panel B), age-adjusted PageRank still predicts market value after controlling for ln(forward citations): in the firm-fixed-effects specification (column 5), its coefficient is positive and significant while the citation coefficient is essentially zero.

PageRank is even more robustly positive for KPST text-based scientific importance. In the initial specifications (Panel C), every PageRank variant enters positively both with and without fixed effects—the raw-PageRank association does not flip the way it does for KPSS—and in the citation horse race (Panel D) the age-adjusted and R_p measures remain positive conditional on ln(forward citations), with the citation coefficient near zero once year and CPC fixed effects are included (a detailed OLS version is in Appendix Table A.9). Network centrality and private appropriability are thus distinct dimensions. Appendix Figure A.17 shows the binned KPSS–PageRank and KPSS–citation relationships (forward cites ∼linear; PageRank curved).

Landmark inventions. We also compare each metric against a curated list of landmark US inventions: we rank the entire patent universe by the metric and compute recall@k and precision@k—respectively, the share of landmarks that fall in the top k of the ranking and the share of the top k that are landmarks (Figure 3). PageRank concentrates the landmarks more tightly in its upper tail than forward citations do: at the top 0.1% of the ranking, raw and age-adjusted PageRank recall roughly twice as many landmark patents as forward citations, while KPSS/KPST—available only for subsets—recall far fewer. This advantage persists within filing cohort (so it is not merely an age artifact) and on the common subset of patents that have KPSS/KPST values (so it is not merely coverage); both diagnostics are in Appendix Section B.

Table 3: External Validation: KPSS Value and KPST Importance Regressions

Panel A. KPSS stock-market value: initial specifications (no citation control).

	With year/CPC FE			No fixed effects		
	raw lnPR	age-adj PR	R_p	raw lnPR	age-adj PR	R_p
	(1)	(2)	(3)	(4)	(5)	(6)
PageRank-family coef.	+0.1773	+0.1391	+0.1095	-0.0512	+0.0798	+0.0841
	(0.0356)	(0.0264)	(0.0212)	(0.0818)	(0.0312)	(0.0225)
Year + CPC FE	Yes	Yes	Yes	No	No	No
R^2	0.0635	0.0645	0.0637	0.0005	0.0015	0.0022

Panel B. KPSS stock-market value: citation horse race.

	cit only	raw PR	age-adj	age-adj	age-adj	R_p
	(1)	(2)	(3)	(4)	(5)	(6)
PageRank-family coef.		-0.3064	-0.1083	+0.0136	+0.0362	+0.0100
		(0.0857)	(0.0559)	(0.0764)	(0.0173)	(0.0216)
ln(citations)	+0.0843	+0.2232	+0.2969	+0.0779	-0.0010	+0.2236
	(0.0456)	(0.0418)	(0.0550)	(0.0761)	(0.0105)	(0.0313)
Fixed effects	None	None	Yr+CPC	None	Yr+CPC+Firm	Yr+CPC
R^2	0.0031	0.0110	0.0772	0.0031	0.8273	0.0759

Panel C. KPST text-importance: initial specifications (no citation control).

	With year/CPC FE			No fixed effects		
	raw lnPR	age-adj PR	R_p	raw lnPR	age-adj PR	R_p
	(1)	(2)	(3)	(4)	(5)	(6)
PageRank-family coef.	+0.0578	+0.0468	+0.0377	+0.0710	+0.0583	+0.0471
	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0005)
Year + CPC FE	Yes	Yes	Yes	No	No	No
R^2	0.5262	0.5264	0.5158	0.0575	0.0520	0.0339

Panel D. KPST text-importance: citation horse race.

	cit only	raw PR	age-adj	age-adj	R_p
	(1)	(2)	(3)	(4)	(5)
PageRank-family coef.		+0.0267	+0.0467	+0.0175	+0.0281
		(0.0002)	(0.0001)	(0.0001)	(0.0003)
ln(citations)	+0.0550	+0.0433	+0.0001	+0.0461	+0.0190
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Fixed effects	None	None	Yr+CPC	None	Yr+CPC
R^2	0.0804	0.0848	0.5264	0.0830	0.5222

Notes: Panels A–B use the KPSS stock-market patent value (Kogan et al., 2017) as dependent variable ($\ln(\xi_{\text{real}})$; $N = 3{,}289{,}136$ in every column), with standard errors clustered by firm (permco). Panels C–D use the KPST text-importance measure (Kelly et al., 2021) as dependent variable ($\ell q_{sim,010}$; $N = 6{,}732{,}002$ in every column), with HC1 robust standard errors; KPST has no firm identifier, so there is no firm-FE column. Standard errors are in parentheses. The PageRank-family coefficient is on raw ln(PageRank), the age-adjusted within-cohort z-score, or the Mariani R_p, as labeled. Dropping fixed effects (Panels A and C, columns 4–6) collapses the raw-lnPR association under KPSS (and even flips its sign), while the age-adjusted and R_p measures stay positive under both benchmarks.

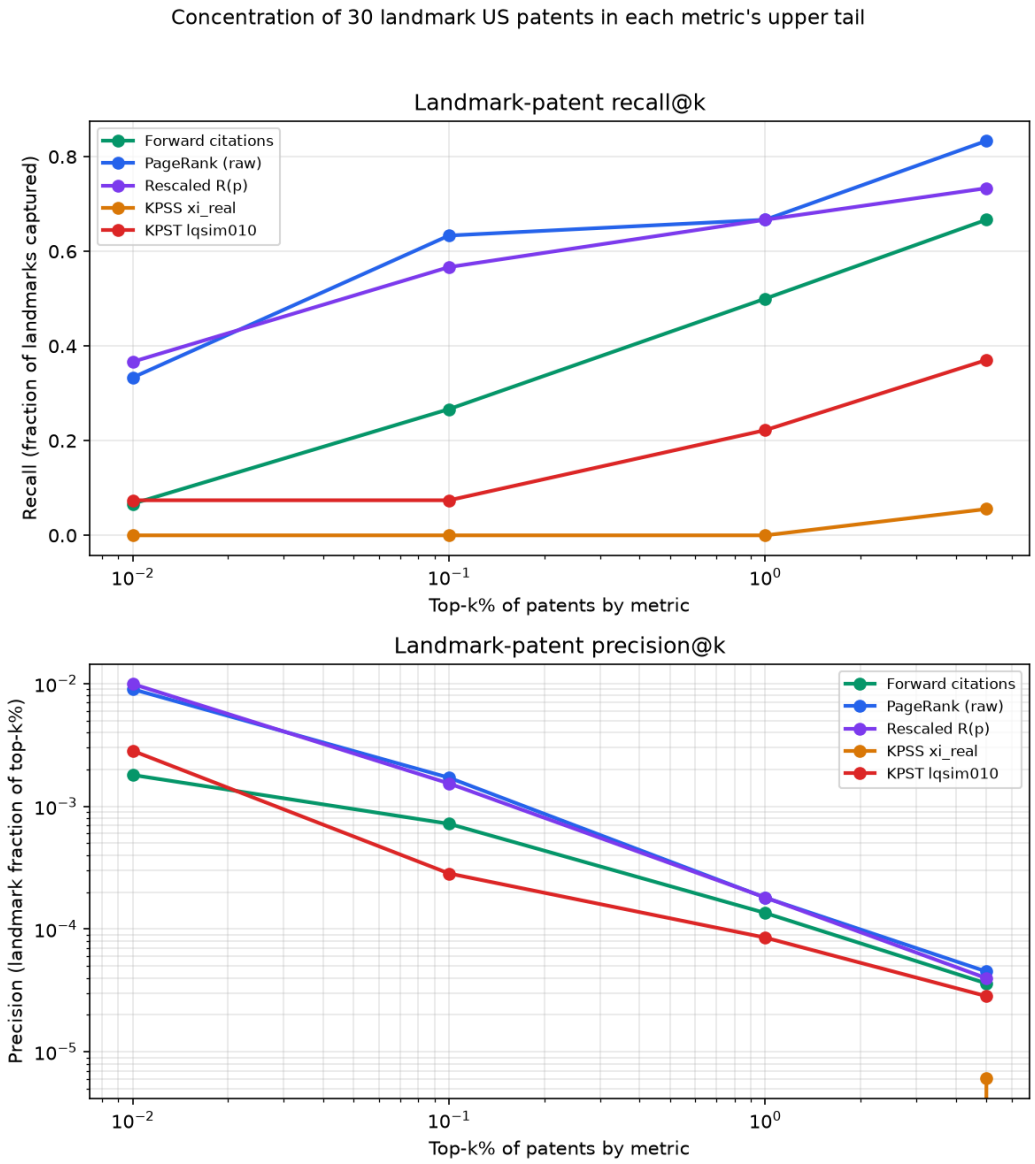


Figure 3: Landmark-Patent Recall and Precision: PageRank vs. Citations vs. Market/Text Measures

Notes: Recall@k (top) and precision@k (bottom) for a curated list of landmark US patents, ranking the universe by each metric. PageRank (raw and age-rescaled) captures landmarks in its upper tail far more efficiently than forward citations; KPSS/KPST are evaluated over their own coverage.

6 Under-Appreciated Patents: What PageRank Sees that Citations Miss

One illustration of what PageRank adds over a raw citation count is the class of patents it recovers from citation obscurity. Because edges run from the citing to the cited patent, a patent accrues PageRank from the patents that cite it; a patent with *few* citations can therefore still attain a high PageRank if the few patents that do cite it are themselves important. These under-appreciated patents—upstream foundations of influential descendants—are invisible to a measure that merely counts citations, yet they are the kind of inventions a value proxy should surface.

The phenomenon. Consider US grants with at most the median number of forward citations (three or fewer). Within this low-citation stratum, the patents that PageRank nonetheless places in its top decile are cited by patents that average the 60^{th} percentile of PageRank, versus the 33^{rd} percentile for the rest of the stratum: their handful of citers are, on average, roughly twice as central.

Table 4 gives examples: each sits in the top tenth of a percent of PageRank but carries only one to three forward citations. They span an early automobile speed indicator (1906), the Baldwin–Monroe calculating machine (1913), Brewster's color-cinematography and color-photography patents (1917–1924), an AT&T picture-transmission (facsimile) patent (1926), and an AT&T enciphering–deciphering (cipher) system from 1946. In each case the lone citation descends from a foundational follow-on invention, which is why PageRank—but not a citation count—registers the parent's importance. (Further extreme cases are in Appendix Table A.2.)

Figure 5 traces four of these foundations to the high-PageRank landmark that cites them: the 1913 Baldwin–Monroe mechanical calculating machine to RCA's 1948 electronic computer, a 1926 AT&T picture-transmission patent to Zworykin's 1952 optical character-recognition device, a 1946 cipher patent to a 1962 stream cipher, and a 1906 speed indicator to a 1950 electronic aircraft tachometer. A single citation from such a descendant is enough for PageRank to register the foundation's importance.

Breakthroughs, not artifacts of network structure. If high PageRank among low-citation patents merely reflected idiosyncratic network structure, it would not align with any external notion of importance. In fact, it aligns with such measures. Holding the citation count fixed, the top-PageRank decile of the low-citation stratum is 1.8 times as likely to be a text-based breakthrough in the sense of Kelly et al. (2021). Figure 6 shows the relationship

Table 4: Under-Appreciated Patents: Few Citations, High PageRank

Patent	Invention (assignee)	Grant	Cites	PR pctile
US-815778	Electrical speed-indicator	1906	1	99.94
US-1080245	Calculating machine (Monroe)	1913	1	99.88
US-1222925	Film for color cinematography (Brewster)	1917	3	99.89
US-1403835	Frequency-control system	1922	1	99.98
US-1508916	Color photography (Brewster)	1924	3	99.90
US-1602469	Electrical transmission of pictures (AT&T)	1926	1	99.93
US-2403888	Enciphering and deciphering system (AT&T)	1946	1	99.97

Notes: US grants in the top tenth of a percent of PageRank that nonetheless carry only one to three forward citations ('Cites'); 'PR pctile' is the global PageRank percentile. Each is an upstream foundation whose lone citation comes from a high-PageRank descendant, so PageRank registers its importance while a raw citation count does not. Titles and assignees verified against Google Patents.

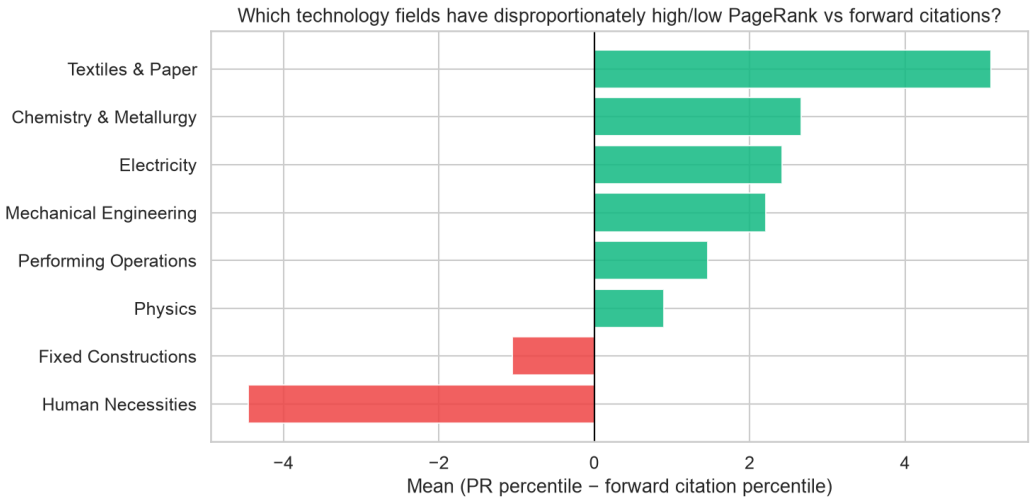


Figure 4: Where PageRank and Forward Citations Diverge, by CPC Section

Notes: Mean percentile divergence (PageRank percentile minus forward-citation percentile) by CPC section (A–H). Positive (green): fields ranked higher by PageRank; negative (red): the reverse.

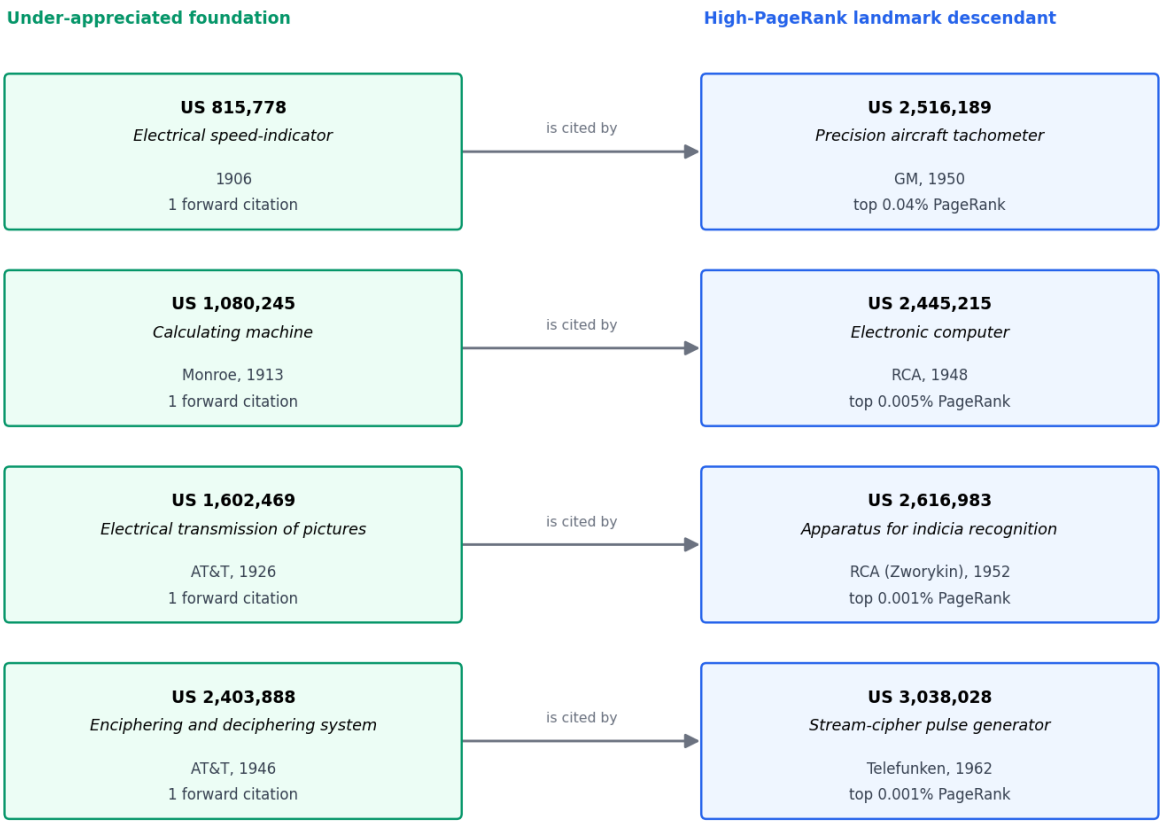


Figure 5: Under-Appreciated Foundations and the Landmark Descendants that Cite Them

Notes: Each low-citation foundation (left) is paired with the highest-PageRank patent that cites it (right), with grant year, assignee, and the descendant's global PageRank percentile. Titles verified against Google Patents.

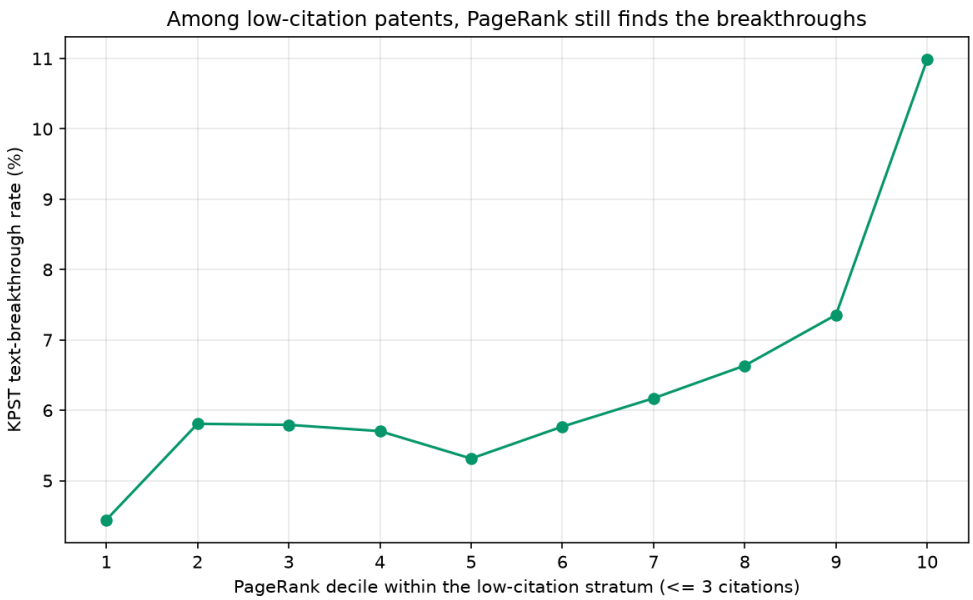


Figure 6: Among Low-Citation Patents, PageRank Still Identifies Breakthroughs

Notes: US grants with at-or-below-median forward citations (three or fewer), binned into deciles of PageRank *within* this low-citation stratum (horizontal axis). The vertical axis is the share that are Kelly et al. (2021) text-based breakthroughs, which rises steeply with PageRank even though the citation count is held roughly fixed. Curated landmarks are heavily cited by construction and so essentially never fall in this low-citation stratum.

is monotone: among patents with three or fewer citations, the text-breakthrough rate climbs from about 4% in the bottom PageRank decile to 11% in the top. PageRank recovers important inventions that a citation count overlooks.

Consistent with the scientific-versus-private distinction drawn in Section 5, these under-appreciated patents do *not* command higher market value—their mean KPSS value is, if anything, lower than that of citation-matched controls. PageRank is recovering scientific and technological foundations rather than private appropriability. This is the margin on which a network measure improves on a citation count, and, unlike the stock-market and text-based measures, it is available for every patent in the corpus.

The disconnect between network centrality and market value is not confined to these low-citation foundations. Table 5 makes the gap concrete at the opposite extreme: several heavily cited, top-PageRank patents—an early non-volatile memory, the Sakoe–Chiba speech-recognition patent, an automotive navigation system—sit in the bottom fraction of a percent of KPSS value, partly because they are foreign-assignee patents whose grant-day US stock reaction KPSS cannot capture.

Table 5: PageRank versus Market Value: High-PageRank Patents the Stock Market Values Low

Patent	Invention (assignee)	Grant	Cites	PR pctile	KPSS pctile
US-4016588	Non-volatile semiconductor memory (NEC)	1977	93	99.99	0.22
US-4158750	Speech recognition, Sakoe–Chiba (NEC)	1979	138	99.93	0.05
US-4194153	Digital radio paging system (NEC)	1980	49	99.99	0.11
US-4459667	Automotive navigation system (Nissan)	1984	61	99.97	0.20
US-4811207	Distributed-database join (Oki)	1989	85	99.91	0.02
US-4870698	Mobile-radio power control (Oki)	1989	93	99.92	0.04
US-5119367	Packet routing for bursty data (Oki)	1992	80	99.91	0.00
US-7049190	ZnO thin-film transistor (Sanyo)	2006	3807	99.93	0.00

Notes: US grants in the top tenth of a percent of PageRank—and, unlike the under-appreciated patents of Section 6, heavily cited ('Cites')—whose Kogan et al. (2017) KPSS market value sits in the bottom quarter of a percent ('KPSS pctile'). Titles and assignees verified against Google Patents. Notably, most are foreign-assignee patents (NEC, Oki, Sanyo, Nissan): KPSS infers value from the grant-day stock-market reaction of the US-listed owner, so it is least informative exactly here, whereas PageRank scores the whole corpus. The divergence thus reflects both the scientific-versus-private-value distinction and the limited coverage of market-based measures.

Technology fields. Where PageRank and forward citations disagree is itself systematic across technology. Figure 4 shows the mean percentile divergence by CPC section: some fields are ranked systematically higher by PageRank than by raw counts (their patents are cited by other high-impact patents), others the reverse (many citations from peripheral work). This field structure motivates the age-and-technology-adjusted variant. Field-level descriptives (mean PageRank by section, generality/originality, the age×field heat map) are in Appendix Section B.

7 Patent Term and Patent Impact

A central promise of a general-purpose value proxy is that it may be used in applied research designs to replace raw or citation-weighted patent counts. We illustrate this by revisiting Abrams (2009)'s question of whether a longer patent term spurs innovation, now weighting patents by PageRank.

The 1995 TRIPS agreement changed US patent term from seventeen years from grant to twenty years from filing. Because pre-TRIPS examination pendency varied across technologies, the *expected* term change differed across patent classes: short-pendency classes gained effective term while long-pendency classes lost it. Following Abrams (2009), we measure each USPC class's expected term extension in days as 1095 − (pre-TRIPS mean pendency)

Table 6: Patent term and the value of innovation

Outer window (mo)	Patent counts (6) (1)	Patent counts (12) (2)	Patent counts† (6) (3)	Cite-weighted (6) (4)	**PR-weighted** (6) (5)	**PR-wtd†** (6) (6)	PR per patent (6) (7)
After × Extension	0.02511**	0.008412	0.04139**	0.02737*	0.02327	0.003788	0.000434
	(0.0103)	(0.008)	(0.0209)	(0.0165)	(0.0193)	(0.0481)	(0.000387)
After	-9.91*	-10.8**	-17.15*	-14.96	-7.557	1.111	-0.1634
	(5.4)	(4.32)	(10.3)	(9.3)	(9.28)	(23.7)	(0.21)
Extension (days)	-0.07591***	-0.06452***	-0.08405***	-0.14***	-0.1564***	-0.1467***	-0.003077***
	(0.0261)	(0.0241)	(0.0285)	(0.03)	(0.0301)	(0.0334)	(0.000547)
CPC-section dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class-specific trends	No	No	Yes	No	No	Yes	No
N (class×month)	1,680	4,200	1,680	1,680	1,680	1,680	1,655
R^2	0.097	0.094	0.151	0.105	0.253	0.295	0.327

Notes: Difference-in-differences replicating Abrams (2009, Tbl. 4) on a USPC-class × filing-month panel around the June 8, 1995 TRIPS patent-term change. *Extension* is the class's expected term change in days (1095− pre-TRIPS mean pendency); *After* marks filing months after the change. Outcomes are monthly patent counts, citation-weighted counts (Abrams's preferred value proxy), and PageRank-weighted counts (this paper's proposal); citation and PageRank weights are normalized by grant-month mean so all outcomes share a common scale. Columns marked † add class-specific linear time trends. "PR per patent" uses the PageRank-weighted mean per patent. All columns include CPC-section dummies and a linear month trend (absorbed by the class trends in † columns). The 2-month inner window drops the three months around the change. Standard errors clustered by USPC class in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

and estimate his difference-in-differences regression at the patent-class × filing-month level,

$$P_{mc} = \alpha + \beta_1 \,\text{After}_m + \beta_2 \,\text{Extension}_c + \beta_3 \,(\text{After}_m \times \text{Extension}_c) + \beta_4 \, t + \gamma_{s(c)} + \varepsilon_{mc}, \quad (5)$$

where After_m marks filing months after the June 1995 change, $\gamma_{s(c)}$ are technology (CPC-section) fixed effects, t is a linear month trend, and standard errors are clustered by class. The coefficient of interest, β_3, asks whether classes with larger expected extensions produced more innovation afterward. As in Abrams (2009) we estimate over symmetric windows around the change, with a two-month inner window that drops the months of the pre-deadline filing spike.

The point for this paper is that the outcome P_{mc} can be *any* value proxy. Table 6 reports patent counts and citation-weighted counts (as in Abrams) alongside PageRank-weighted counts; the citation and PageRank weights are normalized by grant-month mean so that all three outcomes share a common scale and their coefficients are directly comparable.

At Abrams's preferred six-month window, classes with larger expected extensions show a statistically significant post-TRIPS increase in patent counts ($\beta_3 = 0.025$, $p < 0.05$) and

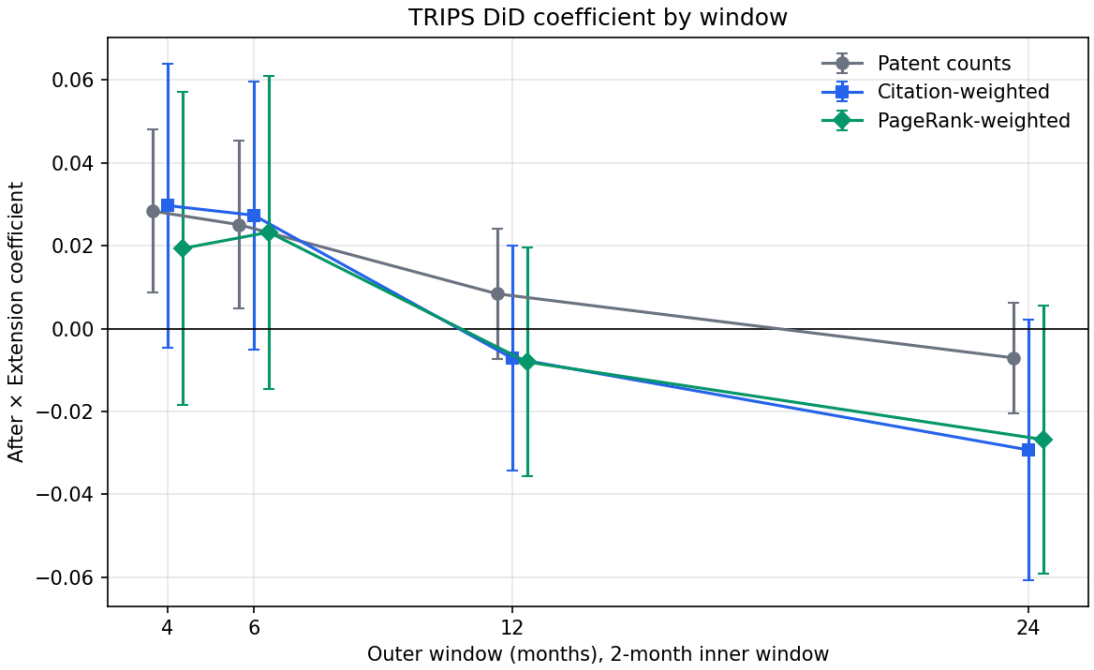


Figure 7: TRIPS Difference-in-Differences Coefficient by Window (Replication of Abrams 2009, Figure 10, with PageRank)

Notes: Each point is the estimated After × Extension coefficient β_3 from Equation (5), by outer window width (months) with a two-month inner window, for patent counts, citation-weighted counts, and PageRank-weighted counts; bars are 95% intervals with standard errors clustered by USPC class.

citation-weighted counts (0.027, $p < 0.1$), reproducing his finding. The PageRank-weighted estimate has the same sign and nearly the same magnitude (0.023) but is not statistically significant: because PageRank is far more concentrated than citation counts, the class-month PageRank totals are noisier and the standard errors correspondingly wider. Figure 7 traces β_3 across window widths. As in Abrams's own Figure 10, the effect is positive at short windows and fades toward—and below—zero as the window widens and longer-run secular trends (the late-1990s surge in long-pendency software and biotechnology patenting) contaminate the cross-class comparison; the short-window estimates are therefore the credible ones.

We build the panel from a fresh pull of day-level filing and grant dates selected on *filing* date, so patents filed near the change are retained with their true grant dates regardless of grant year; this removes the right-truncation that would otherwise bias the wider windows. We present the design as a concrete demonstration that a PageRank-weighted outcome slots directly into an existing patent-term research design—and that it yields conclusions consistent with citation-weighting—rather than as a settled causal estimate.

8 Conclusion

This paper constructs, validates, and studies a family of PageRank-based measures of patent impact for the universe of US granted patents. We use the US-grant-to-US-grant citation graph (12.0 million nodes, 160 million directed edges) and compute PageRank scores together with age-adjusted and age-and-technology-adjusted variants. Each measure weights a received citation by the network centrality of the citing patent, capturing recursive, indirect influence through citation chains beyond what raw forward citation counts can represent.

The empirical analysis documents several systematic patterns. The PageRank distribution is heavily right-skewed, with a power-law tail and extreme concentration. A small number of foundational patents—predominantly in computing, semiconductors, and biotechnology—account for a disproportionate share of total PageRank mass. PageRank and raw citation count are positively correlated but diverge meaningfully at the individual patent level, with the divergence exhibiting systematic patterns by technology field, filing cohort, and assignee type. Age adjustment is essential: raw PageRank overweights older cohorts, while rolling-window rescaling surfaces more recent landmark inventions and provides the appropriate cross-era comparison. The validation results show that PageRank is especially informative for scientific and technological importance, including landmark-patent recall and KPST text-based breakthrough status, while its relationship with private stock-market value is weaker and sign-sensitive. Notably, among patents with few forward citations, those PageRank ranks highly are markedly more likely to be text-based breakthroughs than citation-matched peers—the under-appreciated foundations that a citation count cannot surface.

Several directions could build on this work. On the data side, the public release could be extended to account for applications, foreign patents, and examiner notes. On the measurement side, the gap between PageRank and private market value invites further benchmarking against licensing revenues, renewal and maintenance-fee decisions, and litigation outcomes, which would brighten the line between scientific and private value that the KPSS and KPST comparisons begin to draw. It could also be interesting to explore variants on PageRank that use other measures of downstream influence besides second-order citations, such as the downstream KPSS and KPST scores. Finally, and most generally, our PageRank measures are defined for every patent and can serve as general-purpose impact proxies in downstream research on innovation, knowledge diffusion, and the returns to R&D.

References

Abrams, David S., "Did TRIPS Spur Innovation? An Analysis of Patent Duration and Incentives to Innovate," *University of Pennsylvania Law Review*, 2009, *157* (6), 1613–1647.

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

Ashtor, Jonathan H., "Investigating Cohort Similarity as an Ex Ante Alternative to Patent Forward Citations," *Journal of Empirical Legal Studies*, 2019, *16* (4), 848–880.

Brin, Sergey and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, 1998, *30* (1–7), 107–117.

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

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," Working Paper 8498, National Bureau of Economic Research 2001.

_ , _ , and _ , "Market Value and Patent Citations," *RAND Journal of Economics*, 2005, *36* (1), 16–38.

Harhoff, Dietmar, Francis Narin, Frederic M. Scherer, and Katrin Vopel, "Citation Frequency and the Value of Patented Inventions," *Review of Economics and Statistics*, 1999, *81* (3), 511–515.

Henderson, Rebecca, Adam B. Jaffe, and Manuel Trajtenberg, "Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965–1988," *Review of Economics and Statistics*, 1998, *80* (1), 119–127.

Higham, Kyle, Gaétan de Rassenfosse, and Adam B. Jaffe, "Patent Quality: Towards a Systematic Framework for Analysis and Measurement," *Research Policy*, 2021, *50* (4), 104215.

Jaffe, Adam B. and Gaétan de Rassenfosse, "Patent Citation Data in Social Science Research: Overview and Best Practices," *Journal of the Association for Information Science and Technology*, 2017, *68* (6), 1360–1374.

_ **and Manuel Trajtenberg**, *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, Cambridge, MA: MIT Press, 2002.

_ **,** _ **, and Rebecca Henderson**, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 1993, *108* (3), 577–598.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 2021, *3* (3), 303–320.

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

Lanjouw, Jean O. and Mark Schankerman, "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators," *Economic Journal*, 2004, *114* (495), 441–465.

Mariani, Manuel Sebastian, Matúš Medo, and François Lafond, "Early Identification of Important Patents: Design and Validation of Citation Network Metrics," *Technological Forecasting and Social Change*, 2019, *146*, 644–654.

_ **,** _ **, and Yi-Cheng Zhang**, "Identification of Milestone Papers Through Time-Balanced Network Centrality," *Journal of Informetrics*, 2016, *10* (4), 1207–1223.

Pakes, Ariel, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 1986, *54* (4), 755–784.

Reinstaller, Andreas and Peter Reschenhofer, "Using PageRank in the Analysis of Technological Progress Through Patents: An Illustration for Biotechnological Inventions," *Scientometrics*, 2017, *113* (3), 1407–1438.

Silverberg, Gerald and Bart Verspagen, "The Size Distribution of Innovations Revisited: An Application of Extreme Value Statistics to Citation and Value Measures of Patent Significance," *Journal of Econometrics*, 2007, *139* (2), 318–339.

Trajtenberg, Manuel, "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *RAND Journal of Economics*, 1990, *21* (1), 172–187.

Verspagen, Bart, "Mapping Technological Trajectories as Patent Citation Networks: A Study on the History of Fuel Cell Research," *Advances in Complex Systems*, 2007, *10* (1), 93–115.

Xu, Shuqi, Manuel Sebastian Mariani, Linyuan Lü, and Matúš Medo, "Unbiased Evaluation of Ranking Metrics Reveals Consistent Performance in Science and Technology Citation Data," *Journal of Informetrics*, 2022, *16* (1), 101244.

Younge, Kenneth A. and Jeffrey M. Kuhn, "Patent-to-Patent Similarity: A Vector Space Model," Working Paper, SSRN 2016. SSRN 2709238.

A Supplementary Data and Methodology

This appendix collects supporting detail on the data extraction, patent metadata, citation-data caveats, and the PageRank computation referenced from Sections 3 and 4. All material uses the US-grant-only main specification; full prior-art-graph and robustness variants are in the supplement.

A.1 Extraction and Graph Construction

We extract the full US patent citation graph using a SQL query executed via the BigQuery API. The extraction selects all citation edges where the citing patent has `country_code = 'US'`, creating directed edges from each US citing patent to all patents it cites regardless of the cited patent's country of origin. Each edge thus represents the relation "US patent A lists publication B as prior art." The exact query is reproduced at the end of this subsection.

The result is exported directly from BigQuery to Google Cloud Storage as Snappy-compressed Parquet using BigQuery's `EXPORT DATA` functionality, which avoids materializing the full 353-million-row table in Python memory and produces compact, efficiently compressed files. The Parquet files are then downloaded to a compute cluster for local processing.

We apply the following three cleaning steps to the raw edge list before constructing the citation graph used for PageRank computation:

1. **Null removal.** Edges with null or empty cited publication numbers are dropped. No such edges were found after the `WHERE` filter in the SQL query.

2. **Self-loop removal.** Patents citing themselves are removed (231 cases, representing $6.5 \times 10^{-5}\%$ of all edges).

3. **Duplicate edge collapse.** Multiple citations by the same citing patent to the same cited publication are collapsed to a single directed edge. No duplicates were detected in the current extraction.

The main specification restricts the graph to US granted patents on both sides—edges where a US grant cites another US grant—so the impact measure is defined on a single, homogeneous population. The full prior-art graph, which additionally includes US published applications and foreign cited publications (32.6 million nodes, 353 million edges), is analyzed in the supplement as a robustness check.

The directed US patent citation graph is extracted from Google Patents Public Data on BigQuery with the following query; edges run from each US citing patent to every publication it cites.

```
SELECT
  p.publication_number AS citing_publication_number,
  c.publication_number AS cited_publication_number,
  p.country_code       AS citing_country_code,
  p.filing_date        AS citing_filing_date,
  p.grant_date         AS citing_grant_date
FROM `patents-public-data.patents.publications` AS p,
     UNNEST(p.citation) AS c
WHERE c.publication_number IS NOT NULL
  AND c.publication_number != ''
  AND p.country_code = 'US'
```

A.2 Patent Metadata

For empirical analysis, we augment the PageRank scores with two metadata variables drawn from the same BigQuery table. First, the `filing_date` field in the Google Patents corpus stores dates as eight-digit integers in YYYYMMDD format. We extract the filing year by integer division by 10,000. Filing year is preferred over grant year for measuring when the underlying invention was created, as the filing decision is made by the inventor and is less subject to administrative delays.

Second, we extract the first element of the `cpc` array for each patent, giving the primary Cooperative Patent Classification code assigned to the patent. From this code we derive two aggregations:

- *CPC section* (first letter, A through H): the broadest technology category, corresponding to one of the eight sections listed in Table A.1.
- *CPC group* (first four characters, e.g. `H04L`): a finer classification capturing specific technology subfields.

The metadata merge is performed as a join from the PageRank scores to the BigQuery metadata on publication number. Patents that appear only as citation targets (non-US-prefix publications) may not match the US-focused metadata query and will retain null values for filing year and CPC.

Table A.1: Cooperative Patent Classification (CPC) Sections

Section	Technology area
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons
G	Physics
H	Electricity

Notes: The eight top-level sections of the Cooperative Patent Classification (CPC) system, jointly developed by the USPTO and the European Patent Office.

A.3 Citation Data Considerations

Following Jaffe and de Rassenfosse (2017), we highlight several methodological considerations relevant to the interpretation of our results. First, patent citations in the USPTO record include citations identified by the patent applicant or their attorney and citations added by the patent examiner during prosecution. Research suggests that examiner-added citations are less likely to reflect genuine knowledge flows than applicant-added citations (Jaffe and de Rassenfosse, 2017). The Google Patents corpus does not systematically distinguish between the two sources, so our PageRank scores reflect both types of citations.

Second, we observe only citations made up to the extraction date of June 2026. A patent filed in 2024 has had less than two years to accumulate citations, while one filed in 1980 has had over 45 years. This right-censoring creates a systematic age bias: older patents will tend to have higher raw citation counts, and potentially higher PageRank scores, simply because more of their citation lifetime has been observed. We document this bias empirically in the empirical analysis (Section 5) and note that it is a general feature of citation-based patent impact measures (Hall et al., 2001; Jaffe and de Rassenfosse, 2017).

Third, as documented by Abrams et al. (2013), a non-trivial fraction of citations reflect strategic behavior—particularly for high-value patents where incumbents have strong incentives to erect barriers. Such citations may not represent genuine knowledge flows and will none the less contribute to PageRank scores.

A.4 Dangling Node Correction

A *dangling node* is a node with $k_{\text{out}}(v) = 0$: a patent that appears in the graph as a citation target but has no outgoing citation edges. As noted in Section 3, about 16% of nodes in the US-only graph are dangling. When the random walker reaches a dangling node, it has no outgoing edges to follow. If the walker is simply removed at such a node, the total probability mass in the system decreases over iterations.

The standard correction (Brin and Page, 1998) treats dangling nodes as contributing their mass uniformly to all nodes via the teleportation mechanism. Let $D^{(t)} = \sum_{v:\, k_{\text{out}}(v)=0} r^{(t)}(v)$ be the total mass sitting on dangling nodes at iteration t. The corrected update equation is:

$$r^{(t+1)}(v) = \frac{1-d}{N} + d\left(\frac{D^{(t)}}{N} + \sum_{u:\,(u,v)\in\mathcal{E}} \frac{r^{(t)}(u)}{k_{\text{out}}(u)}\right) \tag{6}$$

This ensures that the probability mass is conserved at every iteration. Because a non-trivial share of nodes are dangling, this correction still has a material effect on the final scores.

A.5 Power Iteration and Convergence

Equation (6) is solved by power iteration. Starting from the uniform distribution $r^{(0)}(v) = 1/N$ for all v, we iterate until the absolute L1 change between successive iterations falls below a tolerance τ:

$$\sum_{v\in\mathcal{V}} \left| r^{(t+1)}(v) - r^{(t)}(v) \right| < \tau \tag{7}$$

The absolute L1 norm is scale-invariant: since the scores always sum to 1, this criterion measures what fraction of total probability mass is still shifting between iterations, regardless of graph size. We use $\tau = 10^{-6}$ (i.e., less than 0.0001% of total mass still shifting).

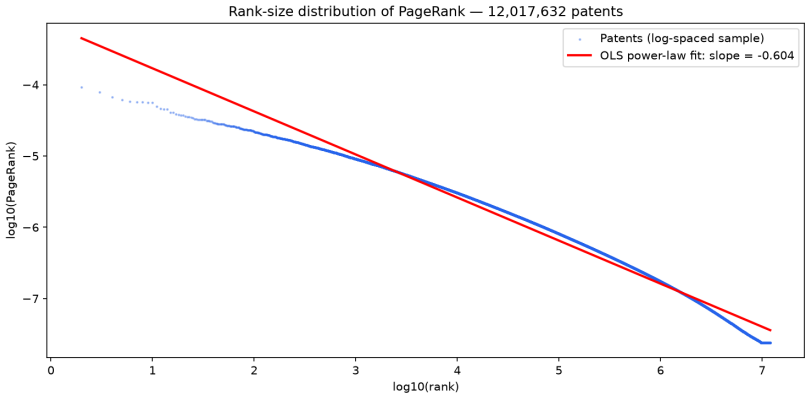
On the US-grant-only graph (12.0 million nodes, 160 million edges), the iteration converges in **29 iterations**, reaching a final abs-L1 delta of 6.45×10^{-7}, confirming convergence to the stationary distribution. The rapid per-iteration progress reflects the highly hierarchical, nearly acyclic structure of the patent citation network.

The computation is implemented in Python using a column-stochastic sparse CSR (Compressed Sparse Row) matrix of the citation graph, with 32-bit floating point entries to reduce memory footprint. The full workflow—reading the US-grant citation edges from Parquet, encoding patent identifiers to integer indices, building the sparse transition matrix, and running 29 power-iteration steps—completes in roughly nine minutes (about 552 seconds) in our benchmark run; the PageRank iterations themselves took 39.6 seconds.

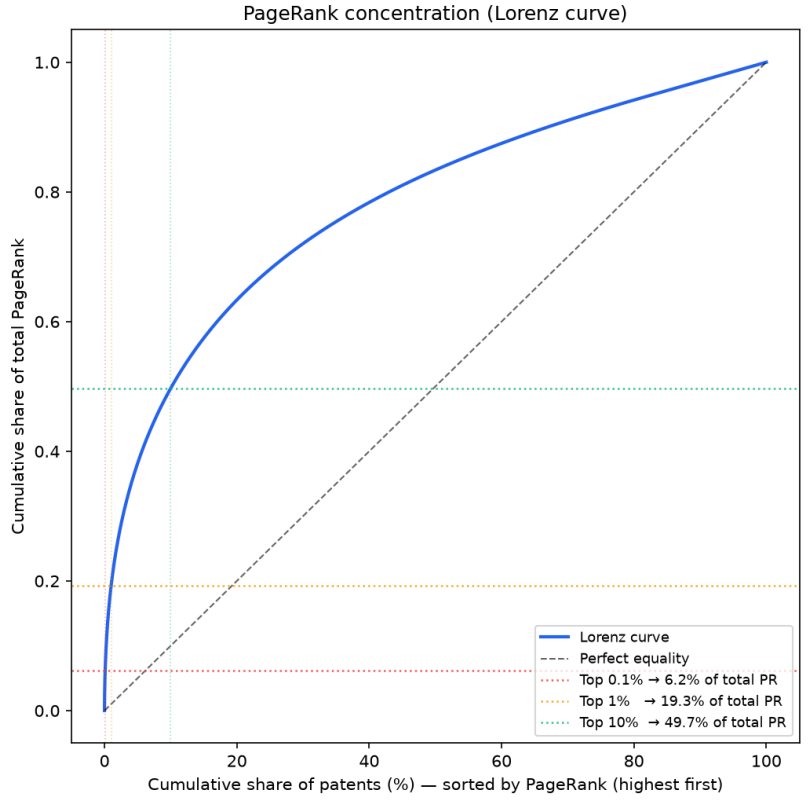
B Additional Exhibits

This appendix collects the supporting tables and figures referenced from the main text. They use the US-grant-only main specification; full prior-art-graph and robustness variants are in the supplement.

B.1 Distribution and Concentration



(a) Rank-size plot (log-log, base 10)



(b) Lorenz curve of PageRank concentration

Figure A.1: Power-Law Structure and Concentration of PageRank

Notes: Left: $\log_{10}$ rank-size plot with an OLS power-law fit. Right: Lorenz curve of cumulative PageRank mass against cumulative share of patents.

B.2 PageRank versus Forward Citations

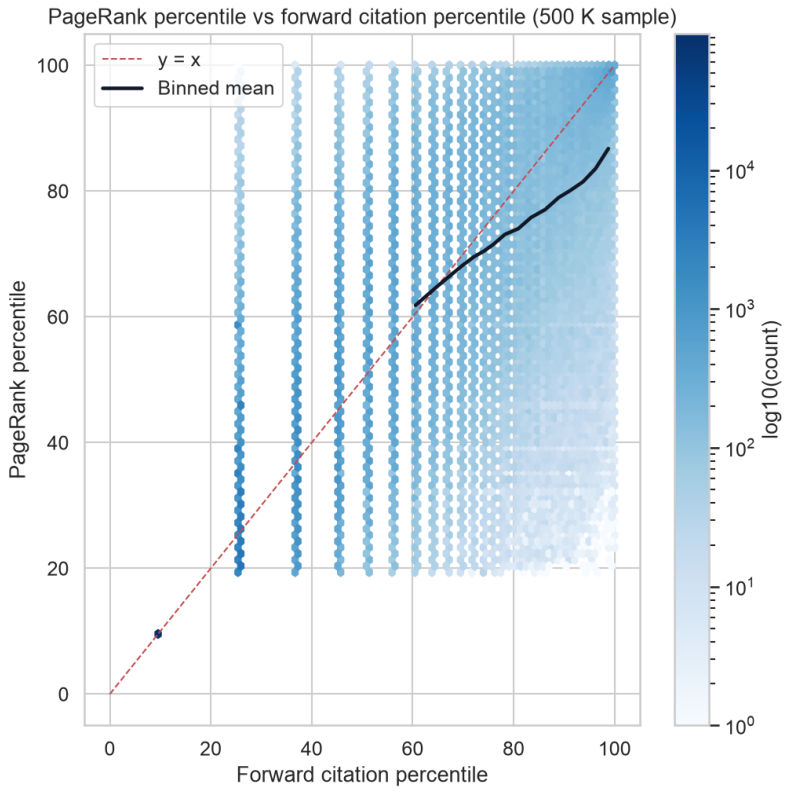


Figure A.2: PageRank Percentile vs. Forward-Citation Percentile

Notes: Hexbin density of PageRank percentile against forward-citation percentile (500K sample) with the 45-degree line and a binned-mean trend.

Table A.2: Extreme PageRank–Citation Percentile Divergence

Panel	Patent	Common name	Product	Grant	c	Δ_v
PageRank $\gg$ *citations*						
	US-3925601	Data switch line circuit		1975	1	+74.09
	US-2786891	Pulse switching system		1957	1	+74.08
	US-3267389	Tunnel-injection amplifier		1966	1	+74.08
	US-1403835	Frequency control system		1922	1	+74.07
	US-3842092	Antibiotic compound		1974	1	+74.07
Citations $\gg$ *PageRank*						
	US-11426160	Smart sensor signal processing		2022	150	−78.77
	US-10881339	Sensor failure detection		2021	149	−78.76
	US-11439470	Robot surgical instrument	da Vinci Surgical System	2022	136	−78.76
	US-11510673	Powered stapler retraction	surgical stapler	2022	139	−78.75
	US-10709495	Motorized RF device bailout		2020	139	−78.75

Notes: Positive $\Delta_v = P_{PR} - P_{cit}$ means PageRank ranks the grant higher than citations; negative values mean the reverse. Restricted to US grants with an observed grant year. The name column shows a short general-audience description; the Product column names a product type or specific product the patent contributed to where one is well established, otherwise blank. Source: `49_divergence.csv`.

Table A.3: Top Patents by PageRank–Citation Percentile Divergence Within Each Grant Decade, 1900–2020

Rank	Patent	Subject	Grant	c	Δ_v
1900s					
1	815778	Speed indicator	1906	1	+74.0
2	714042	Indigo dye making	1902	1	+73.7
3	803734	Gas stove	1905	1	+73.6
4	936622	Coil support	1909	1	+73.6
1910s					
1	1080245	Calculating-machine	1913	1	+74.0
2	1095472	Menthol inhaler	1914	1	+73.8
3	1118228	Oscillation detector	1914	1	+73.6
4	987220	Burglary alarm	1911	1	+73.5
1920s					
1	1403835	Frequency control system	1922	1	+74.1
2	1664453	Morse multiplex telegraphy	1928	1	+74.0
3	1602469	Fax transmission	1926	1	+74.0
4	1581161	Means for reducing interference and obtaining...	1926	1	+74.0
1930s					
1	1904276	Dry rectifier	1933	1	+74.0
2	1845135	Torpedo stop valve	1932	1	+73.9
3	2084414	Electric totalizer	1937	1	+73.9
4	1819197	Early television projector	1931	1	+73.8
1940s					
1	2403888	Enigma-style encryption	1946	1	+74.1
2	2249370	Blood pressure recording	1941	1	+74.1
3	2486360	Supervisory control signals	1949	1	+74.0
4	2291459	Loudspeaker	1942	1	+73.9
1950s					
1	2786891	Pulse switching system	1957	1	+74.1
2	2717892	Streptomycin purification	1955	1	+74.1
3	2876355	Waveform compensation	1959	1	+74.0
4	2843064	Trash burner cover	1958	1	+74.0
1960s					
1	3267389	Tunnel-injection amplifier	1966	1	+74.1
2	3459896	Telephone code calling	1969	1	+74.1
3	3349770	Gynecological prosthesis	1967	1	+74.1
4	2966303	Calculator	1960	1	+74.0
1970s					
1	3925601	Data switch line circuit	1975	1	+74.1
2	3842092	Antibiotic compound	1974	1	+74.1
3	4041023	Labelled peptides	1977	1	+73.9
4	3551296	Malate dehydrogenase test	1970	1	+73.8
1980s					
1	4200839	Cable television broadcasting	1980	1	+73.9
2	4411079	Suction dredging system	1983	1	+72.6
3	4762790	Microplant propagation	1988	1	+72.4
4	4801326	Pollen growth inhibitor	1989	1	+72.3
1990s					
1	5625133	Inbred maize line	1997	1	+73.6
2	5121481	Right-justified word processor display	1992	1	+73.3
3	5182268	Daunorubicin drug composition	1993	1	+71.8
4	5860039	Copier exposure control	1999	1	+71.5
2000s					
1	7522355	Camera lens unit	2009	1	+69.5
2	6362404	Inbred corn seed	2002	1	+68.0
3	6429361	Soybean plant variety	2002	1	+67.7
4	6551902	Semiconductor isolation region	2003	1	+67.5
2010s					
1	9098539	Database token access	2015	1	+72.9
2	8730657	Mobile computing device	2014	1	+70.6
3	9552127	Foldable display control	2017	1	+70.5
4	8359060	Wireless power control	2013	1	+70.4
2020s					
1	10804807	Zero-voltage switching circuit	2020	1	+65.2

Table A.4: Most Under-Cited Important Patents by Grant Decade: Largest PageRank–Citation Divergence Among Top-1% PageRank Patents, 1900–2020

Rank	Patent	Subject	Grant	c	Δ_v
1900s					
1	815778	Speed indicator	1906	1	+74.0
2	714042	Indigo dye making	1902	1	+73.7
3	803734	Gas stove	1905	1	+73.6
4	936622	Coil support	1909	1	+73.6
1910s					
1	1080245	Calculating-machine	1913	1	+74.0
2	1095472	Menthol inhaler	1914	1	+73.8
3	1118228	Oscillation detector	1914	1	+73.6
4	987220	Burglary alarm	1911	1	+73.5
1920s					
1	1403835	Frequency control system	1922	1	+74.1
2	1664453	Morse multiplex telegraphy	1928	1	+74.0
3	1602469	Fax transmission	1926	1	+74.0
4	1581161	Means for reducing interference and obtaining…	1926	1	+74.0
1930s					
1	1904276	Dry rectifier	1933	1	+74.0
2	1845135	Torpedo stop valve	1932	1	+73.9
3	2084414	Electric totalizer	1937	1	+73.9
4	1819197	Early television projector	1931	1	+73.8
1940s					
1	2403888	Enigma-style encryption	1946	1	+74.1
2	2249370	Blood pressure recording	1941	1	+74.1
3	2486360	Supervisory control signals	1949	1	+74.0
4	2291459	Loudspeaker	1942	1	+73.9
1950s					
1	2786891	Pulse switching system	1957	1	+74.1
2	2717892	Streptomycin purification	1955	1	+74.1
3	2876355	Waveform compensation	1959	1	+74.0
4	2843064	Trash burner cover	1958	1	+74.0
1960s					
1	3267389	Tunnel-injection amplifier	1966	1	+74.1
2	3459896	Telephone code calling	1969	1	+74.1
3	3349770	Gynecological prosthesis	1967	1	+74.1
4	2966303	Calculator	1960	1	+74.0
1970s					
1	3925601	Data switch line circuit	1975	1	+74.1
2	3842092	Antibiotic compound	1974	1	+74.1
3	4041023	Labelled peptides	1977	1	+73.9
4	3551296	Malate dehydrogenase test	1970	1	+73.8
1980s					
1	4200839	Cable television broadcasting	1980	1	+73.9
2	4520106	Tartrate breakdown gene	1985	2	+62.6
3	4378418	Photoreceptor charge injection contact	1983	2	+62.5
4	4245341	Encrypted information transmission	1981	2	+62.4
1990s					
1	5625133	Inbred maize line	1997	1	+73.6
2	5121481	Right-justified word processor display	1992	1	+73.3
3	5457275	Inbred corn seed	1995	2	+62.5
4	4947385	Star bus monitoring	1990	2	+62.4
2000s					
1	6275716	High-temperature superconducting material	2001	2	+62.1
2	7045815	Low-leakage semiconductor structure	2006	6	+39.2
3	6526988	Hair reshaping treatment	2003	6	+38.5
4	6599824	Printed microchip wiring	2003	7	+35.3
2010s					
1	7895349	Enterprise messaging protocol	2011	2	+62.1
2	8315440	Animal iris identification	2012	4	+47.9
3	8654304	Flat-panel display	2014	5	+43.2
4	8112804	Cancer detection method	2012	6	+38.9
2020s					
1	10880895	5G downlink control signaling	2020	40	+5.7

B.3 Temporal Patterns and Age Adjustment

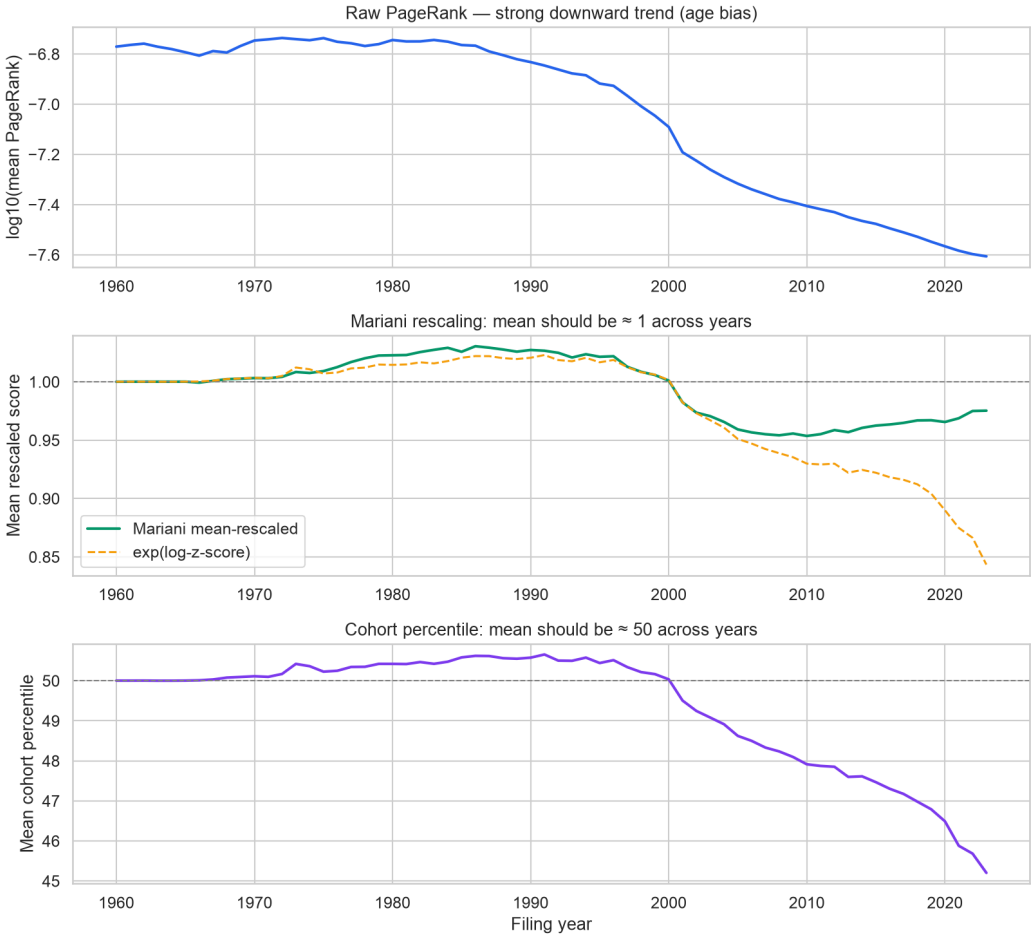


Figure A.3: Age Bias in Raw PageRank and After Cohort Normalization

Notes: Top: mean $\log_{10}$(PageRank) by filing year (raw), showing systematic age bias. Middle: mean cohort-mean-rescaled PageRank ($\tilde{r}_{\text{mean}}$), ≈ 1 when age bias is removed. Bottom: mean cohort percentile, ≈ 50 by construction. Cohorts with fewer than 20 patents excluded.

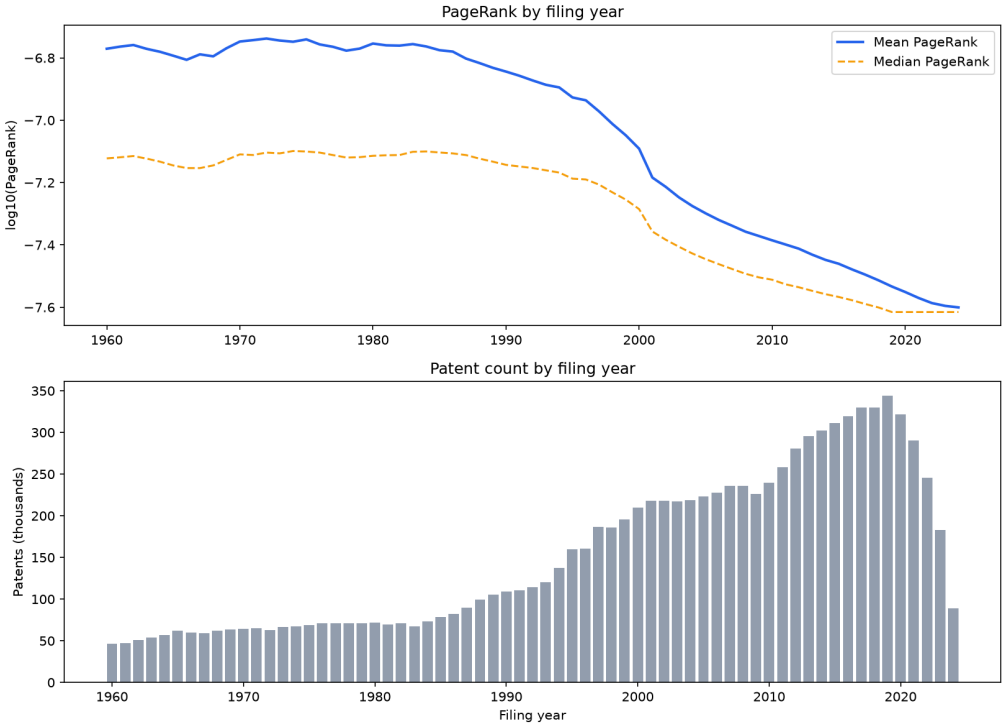
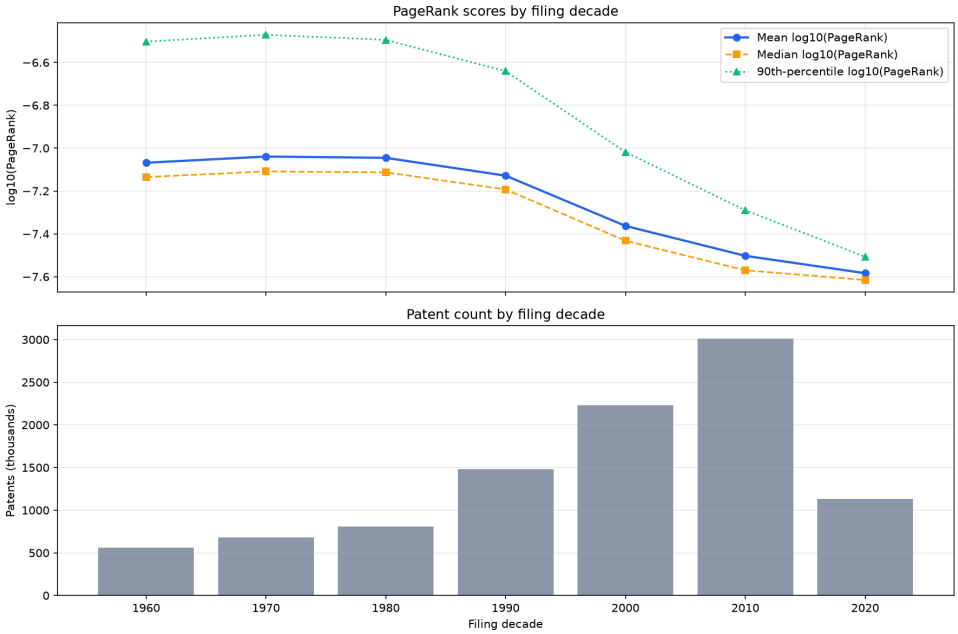
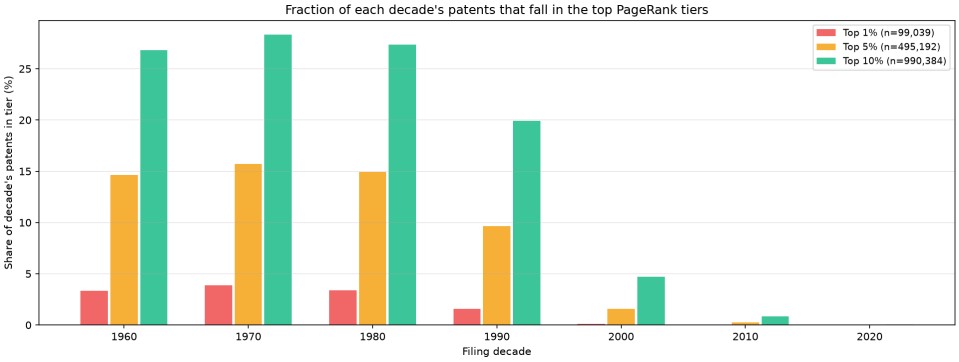


Figure A.4: PageRank by Filing Year

Notes: Mean and median $\log_{10}$(PageRank) by filing year, 1960–2024. The downward trend for recent cohorts reflects citation right-censoring.



(a) PageRank distribution by filing decade



(b) Share of top-k% patents by filing decade

Figure A.5: Patent Impact by Filing Cohort (Decade)

Notes: Older cohorts are over-represented in high-impact bins (cumulative advantage plus right-censoring). Decades shown from 1960.

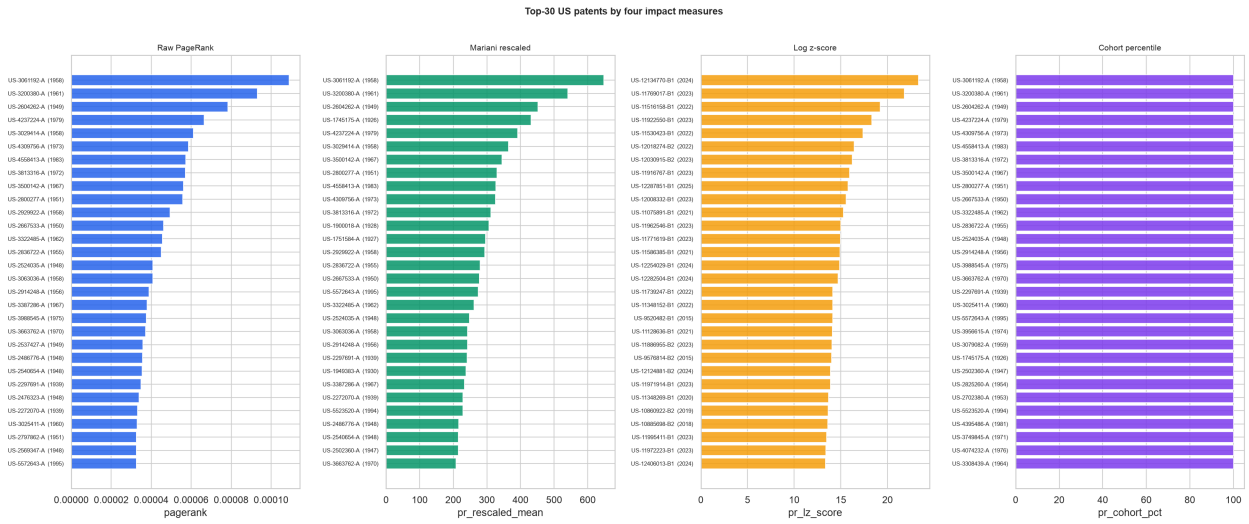


Figure A.6: Top-30 Patents Under Raw vs. Age-Adjusted PageRank Measures

Notes: Top-30 patents by raw PageRank, Mariani cohort-mean rescaling, log z-score, and cohort percentile. Filing year annotated.

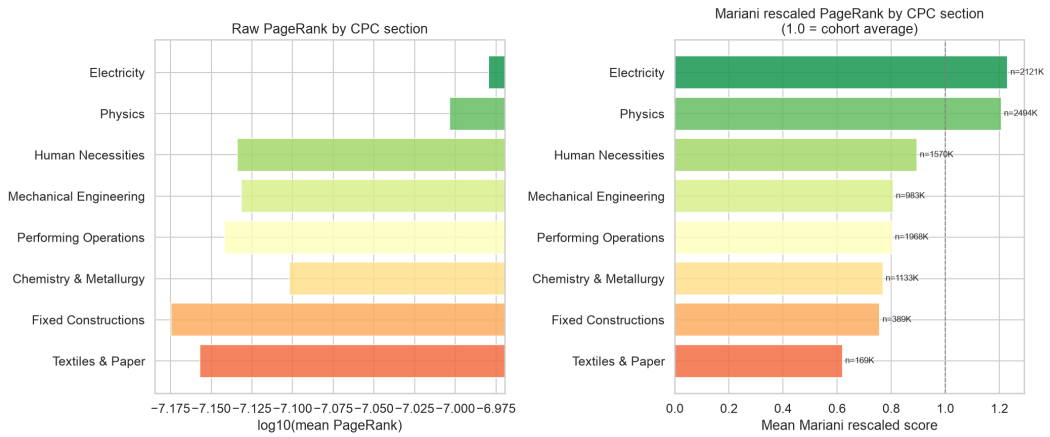


Figure A.7: Age-Adjusted PageRank by CPC Technology Section

Notes: Left: mean $\log_{10}$(PageRank) by CPC section (axis clipped). Right: mean cohort-mean-rescaled PageRank; values above 1 indicate above-cohort-average impact. Sections A–H.

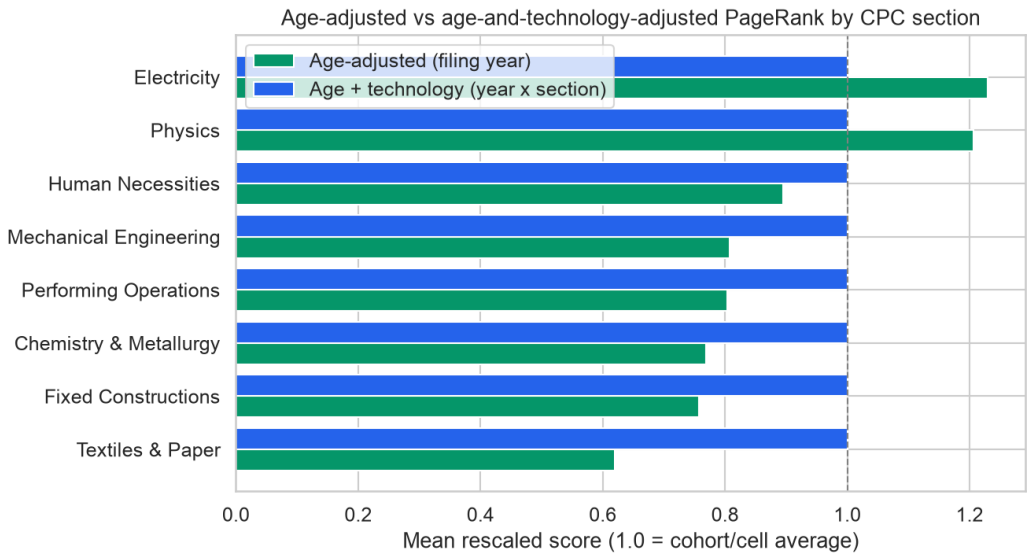


Figure A.8: Age-Adjusted vs. Age-and-Technology-Adjusted PageRank by CPC Section

Notes: Age-adjusted (filing-year) vs. age-and-technology-adjusted (filing-year × CPC-section) mean rescaled PageRank. The age+tech score flattens to ≈ 1 across sections by construction, removing cross-field differences in citing intensity.

B.4 Comparison with Mariani et al. (2019)

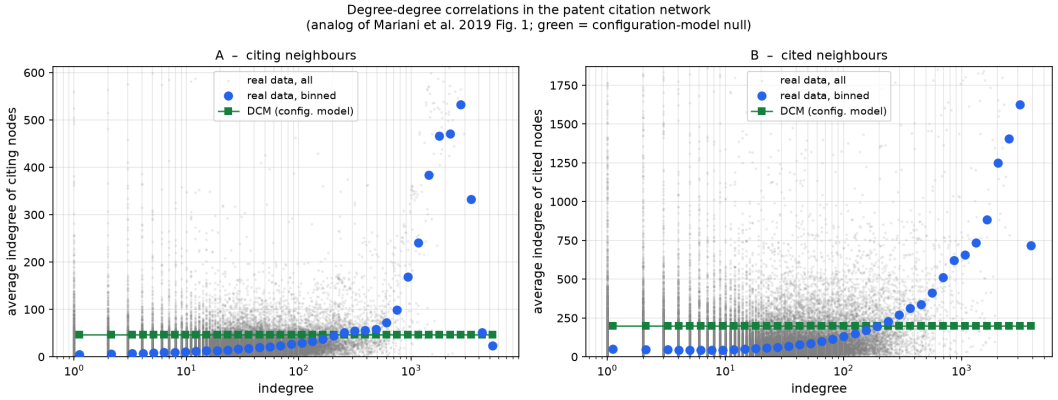


Figure A.9: Degree–Degree Correlations (after Mariani et al. 2019, Fig. 1)

Notes: Gray: raw data; blue: binned means; green: degree-preserving configuration-model null. Data above the null indicate positive assortativity.

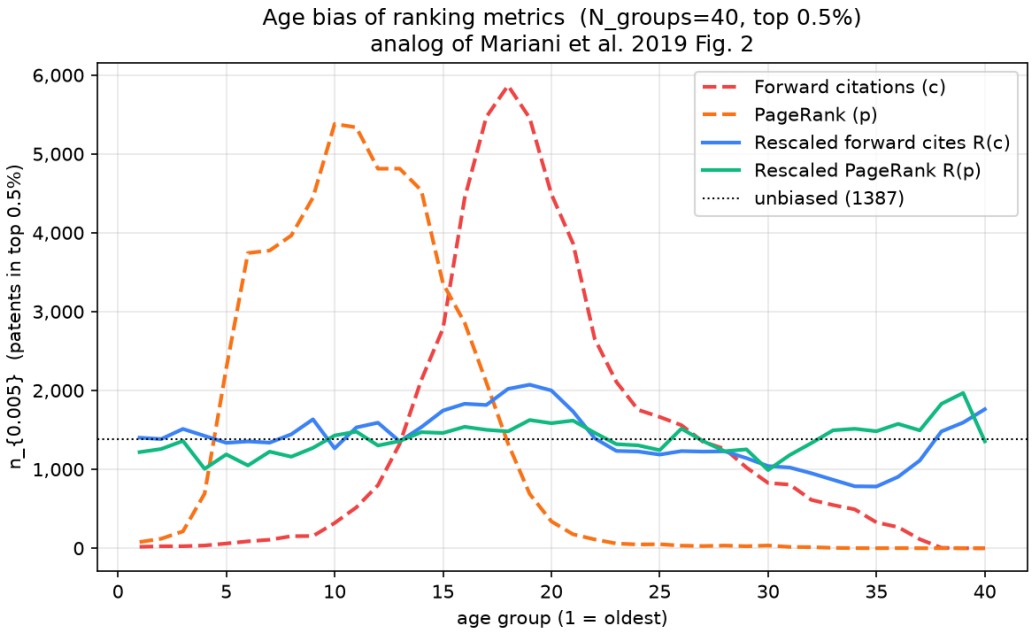


Figure A.10: Age Bias of Ranking Metrics (after Mariani et al. 2019, Fig. 2)

Notes: Patents in 40 equal age groups (1 = oldest); each series is the count in the global top 0.5% by that metric. Raw c and p concentrate in particular cohorts; the age-rescaled $R(c)$, $R(p)$ are roughly flat.

Table A.5: Top-15 US Patents by Raw PageRank p (Analog of Mariani et al. 2019, Table 3)

Rank	Patent	Name	Grant	c	$p \times 10^5$
1	3061192	Early transistorized computer	1962	158	10.90
2	3200380	Computer data processing system	1965	88	9.31
3	2604262	Multiplying and dividing machine	1952	98	7.84
4	4237224	Recombinant DNA	1980	326	6.65
5	3029414	Information handling device	1962	98	6.12
6	4309756	Automatic code compilation	1982	30	5.86
7	4558413	Source code version control	1985	1205	5.73
8	3813316	Microbes with degradative plasmids	1974	68	5.71
9	3500142	Floating-gate MOS transistor	1970	78	5.62
10	2800277	Computer control circuits	1957	37	5.58
11	2929922	Maser	1960	123	4.95
12	2667533	Automatic message switching	1954	46	4.62
13	3322485	Liquid-crystal display element	1967	79	4.57
14	2836722	Atomic clock	1958	26	4.50
15	2524035	First transistor	1950	132	4.09

Notes: Top-15 US patents by raw PageRank score p (scaled by 10^5). The Name column gives the popular product/technique name where one is known, otherwise the patent title; c is the forward-citation count. Most were granted before 1980, illustrating the age bias of raw PageRank. Source: `40_mariani_top15_by_pagerank.csv`.

Table A.6: Top-15 US Patents by Rescaled PageRank $R(p)$ (Analog of Mariani et al. 2019, Table 4)

Rank	Patent	Name	Grant	c	$R(p)$
1	3200380	Computer data processing system	1965	88	109.2
2	3061192	Early transistorized computer	1962	158	106.1
3	12552340	Hydraulic valve safety check	2026	1	105.0
4	1900018	Electric current control	1933	51	103.4
5	9576814	Spacer patterning for chips	2017	3068	102.5
6	4237224	Recombinant DNA	1980	326	101.9
7	2604262	Multiplying and dividing machine	1952	98	100.0
8	12542224	Downhole wireline cables	2026	1	96.7
9	9236267	FinFET cut-mask patterning	2016	3085	95.8
10	5572643	Web browser linking display	1996	2149	94.9
11	9520482	Metal gate cutting	2016	3117	94.6
12	4558413	Source code version control	1985	1205	92.4
13	2297691	Xerography	1942	749	91.5
14	2929922	Maser	1960	123	90.3
15	3813316	Microbes with degradative plasmids	1974	68	89.0

Notes: Top-15 US patents by the rolling-window rescaled PageRank $R(p)$. The Name column gives the popular product/technique name where one is known, otherwise the patent title. Relative to the raw ranking, the age correction lifts recent semiconductor and computing patents into the top ranks. Source: `41_mariani_top15_by_rescaled_pagerank.csv`.

Table A.7: Top Patents by Rescaled PageRank $R(p)$ Within Each Grant Decade, 1900–2020

Rank	Patent	Title	Grant	c	$R(p)$
1900s					
1	925988	Tantalum hardening	1909	4	79.1
2	815778	Speed indicator	1906	1	56.4
1910s					
1	1214022	Wireless telegraphy apparatus	1917	2	81.6
2	960300	Wire stitching tool	1910	52	75.8
1920s					
1	1647474	Variable path routing	1927	67	67.3
2	1718993	Electrical wiring panel	1929	36	65.7
1930s					
1	1900018	Electric current control	1933	51	103.4
2	1949383	Electronic device	1934	26	87.7
1940s					
1	2297691	Xerography	1942	749	91.5
2	2486776	Self-biased electric tube	1949	83	83.5
1950s					
1	2604262	Multiplying and dividing machine	1952	98	100.0
2	2836722	Atomic clock	1958	26	82.2
1960s					
1	3200380	Computer data processing system	1965	88	109.2
2	3061192	Early transistorized computer	1962	158	106.1
1970s					
1	3813316	Microbes with degradative plasmids	1974	68	89.0
2	3500142	Floating-gate MOS transistor	1970	78	85.5
1980s					
1	4237224	Recombinant DNA	1980	326	101.9
2	4558413	Source code version control	1985	1205	92.4
1990s					
1	5572643	Web browser linking display	1996	2149	94.9
2	5523520	Petunia dwarfism gene	1996	5006	82.6
2000s					
1	7356679	Image capture and deployment	2008	957	71.5
2	7365920	Camera lens assembly	2008	82	70.5
2010s					
1	9576814	Spacer patterning for chips	2017	3068	102.5
2	9236267	FinFET cut-mask patterning	2016	3085	95.8
2020s					
1	12552340	Hydraulic valve safety check	2026	1	105.0
2	12542224	Downhole wireline cables	2026	1	96.7

Notes: Top-2 US patents by the rolling-window rescaled PageRank $R(p)$ within each grant decade from 1900 through 2020, ordered by $R(p)$ within decade. c is the forward-citation count; titles are truncated for display. The full source table (top-20 in every decade from 1840 onward) is available as `44_top20_by_era.csv`. The headline display starts in 1900 to avoid sparse nineteenth-century cohorts and poor early metadata.

B.5 Technology Fields

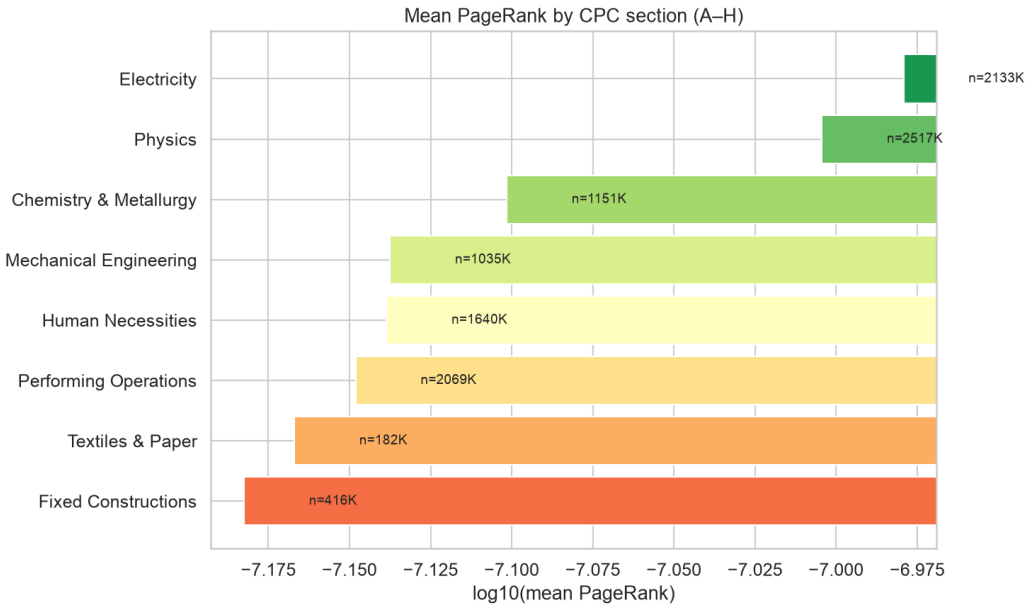


Figure A.11: Mean PageRank by CPC Technology Section

Notes: Mean $\log_{10}$(PageRank) by CPC section (A–H), x-axis clipped to the data range.

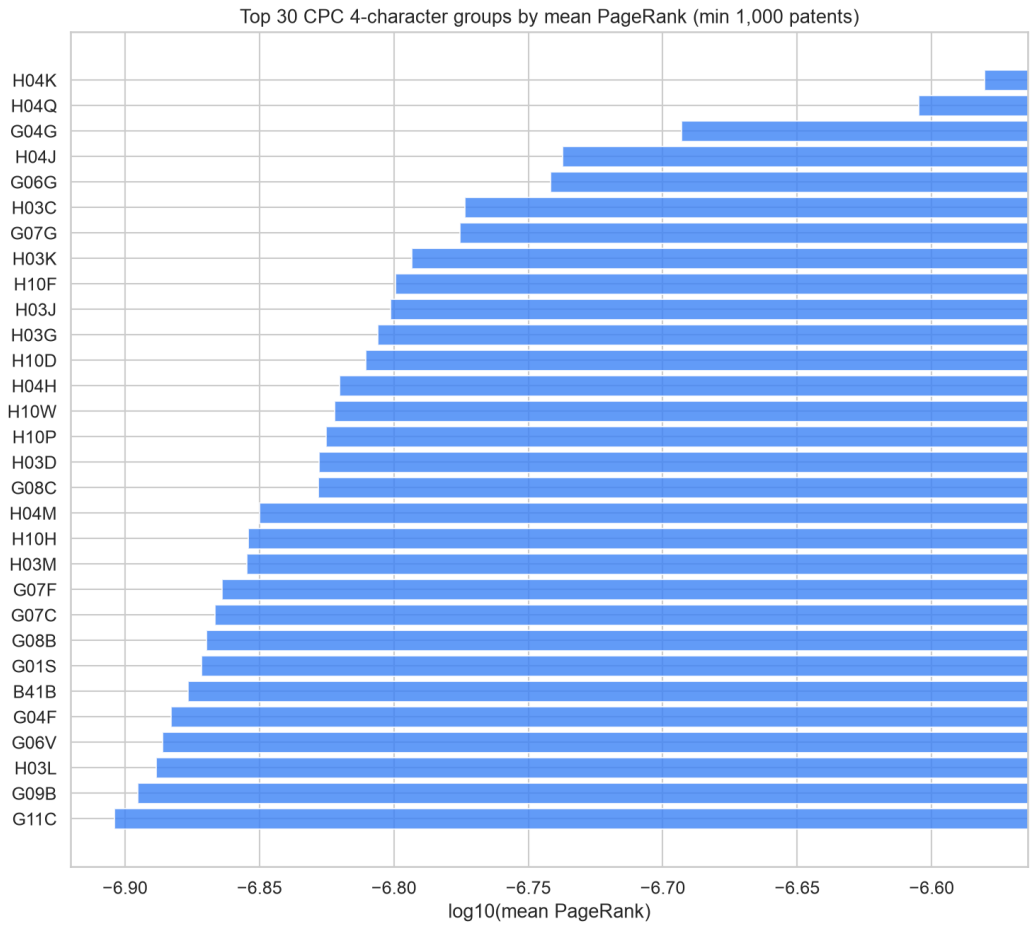
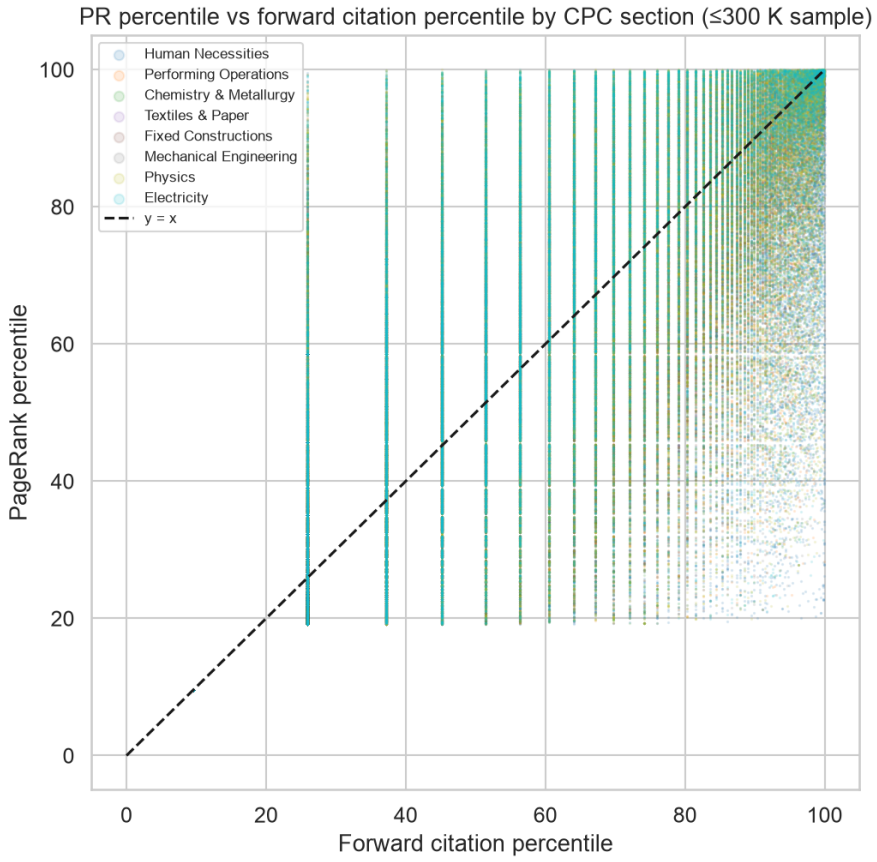
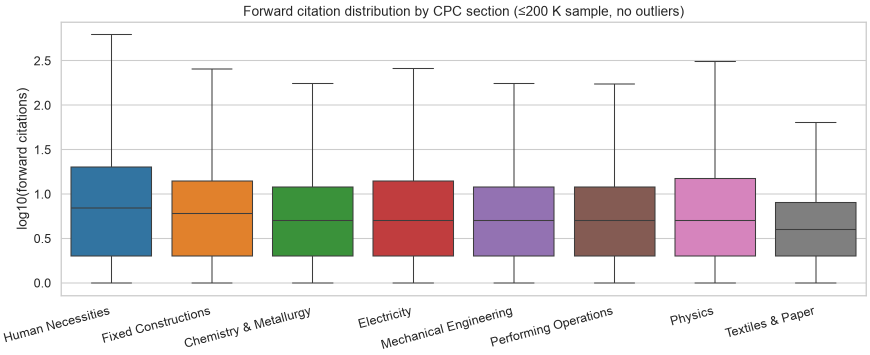


Figure A.12: Top-30 CPC Groups by Mean PageRank

Notes: Mean $\log_{10}$(PageRank) for the top-30 CPC 4-character groups with at least 1,000 patents.



(a) PR vs. citation percentile by CPC section



(b) Citation distribution by CPC section

Figure A.13: PageRank and Citation Distributions by CPC Section

Notes: Left: PageRank vs. citation percentile (300K sample), colored by section. Right: $\log_{10}$(forward citations) by section.

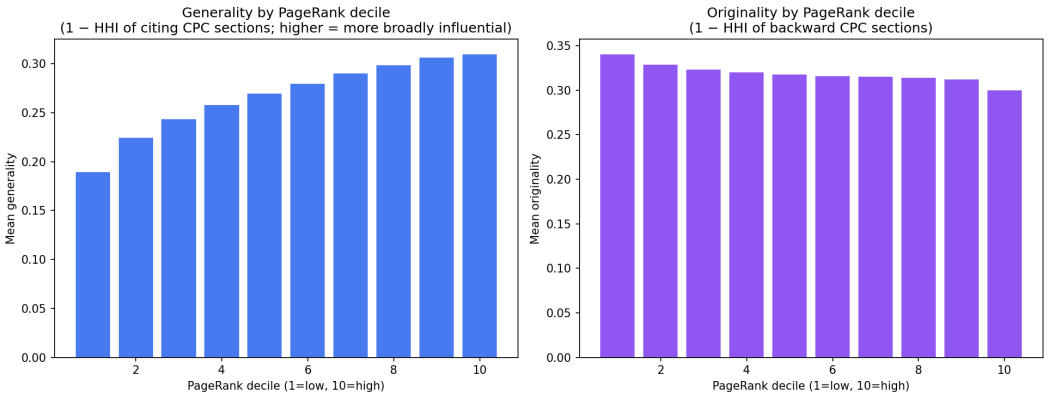


Figure A.14: Generality and Originality by PageRank Decile

Notes: Mean HJT generality (left) and originality (right) by PageRank decile. Computed on the full prior-art citation graph.

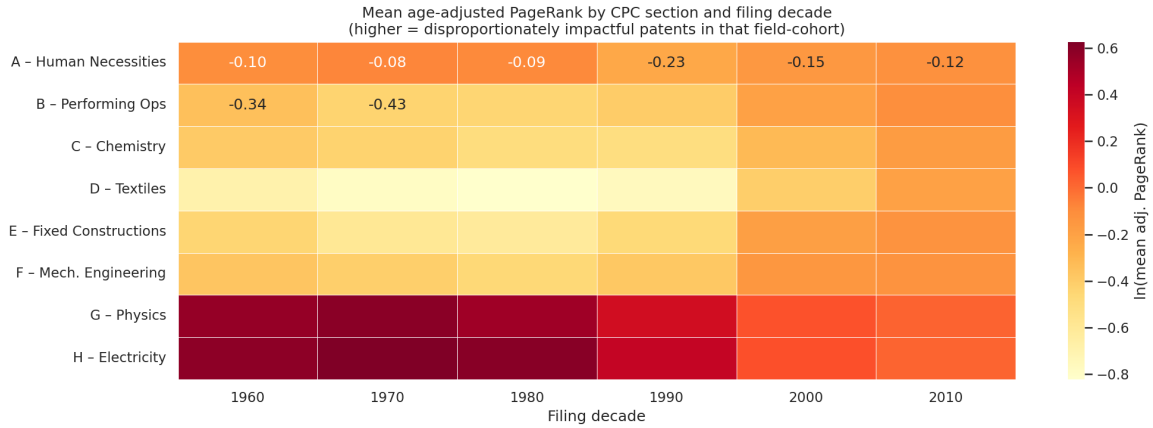
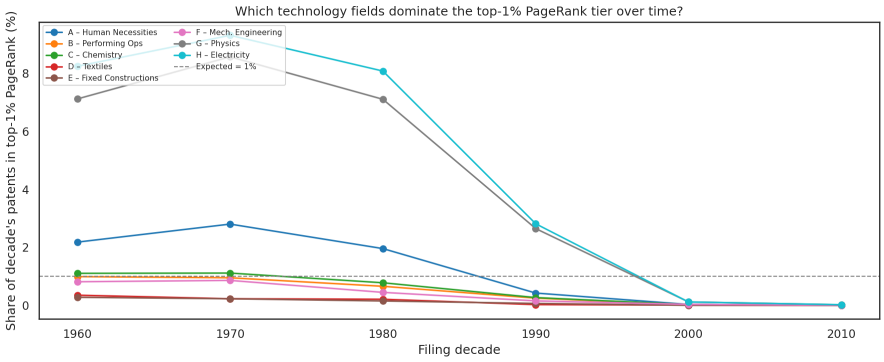
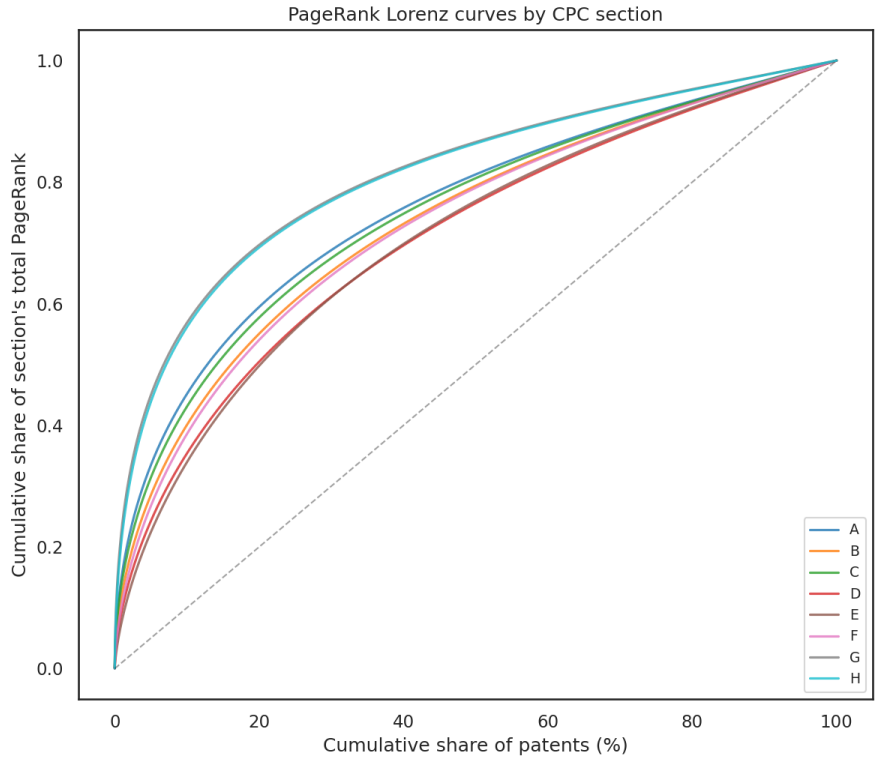


Figure A.15: Mean Age-Adjusted PageRank by CPC Section and Filing Decade

Notes: Each cell is mean cohort-mean-rescaled PageRank for a CPC section × filing decade; values above 1 indicate above-cohort impact.



(a) Top-1% share over time by CPC section



(b) Within-section PageRank Lorenz curves

Figure A.16: Industry Dynamics of Patent Impact

Notes: Left: share of each decade's patents (by section) in the top 1% of the overall PageRank distribution. Right: within-section Lorenz curves.

B.6 External Validation: Detailed Regressions

Additional initial-specification and time-aware-variant tables are available in the project outputs and the supplement.

Table A.8: OLS Regressions: KPSS Patent Value on PageRank and Citation Count

	(1) (cit. only)	(2) (raw PR)	(3) (age-adj. PR)	(4) (+ firm FE)	(5) (Mariani R_p)
ln(citations)	+0.0798	+0.1911	+0.2579	−0.0167	+0.2105
	(0.0513)	(0.0399)	(0.0475)	(0.0120)	(0.0291)
ln(PageRank) (raw)	—	−0.2433	—	—	—
		(0.0954)			
PageRank (age-adj., $\tilde{r}$)	—	—	−0.0492	+0.0614	—
			(0.0436)	(0.0196)	
PageRank (Mariani R_p)	—	—	—	—	+0.0368
					(0.0153)
Filing-year FE	No	No	Yes	Yes	Yes
CPC-section FE	No	No	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No
N	3,375,672	3,375,672	3,375,672	3,375,672	3,375,672
R^2	0.0030	0.0080	0.0711	0.8317	0.0712

Notes: Dependent variable: $\ln(\xi_{real})$, the KPSS stock-market estimate of private patent value (Kogan et al., 2017). Sample: US patents matched to CRSP firms with $\xi_{real} > 0$. "Age-adj. PR" is the within-cohort standardized log-PageRank. Fixed effects by Frisch–Waugh–Lovell demeaning; total R^2 reported. SEs clustered by firm.

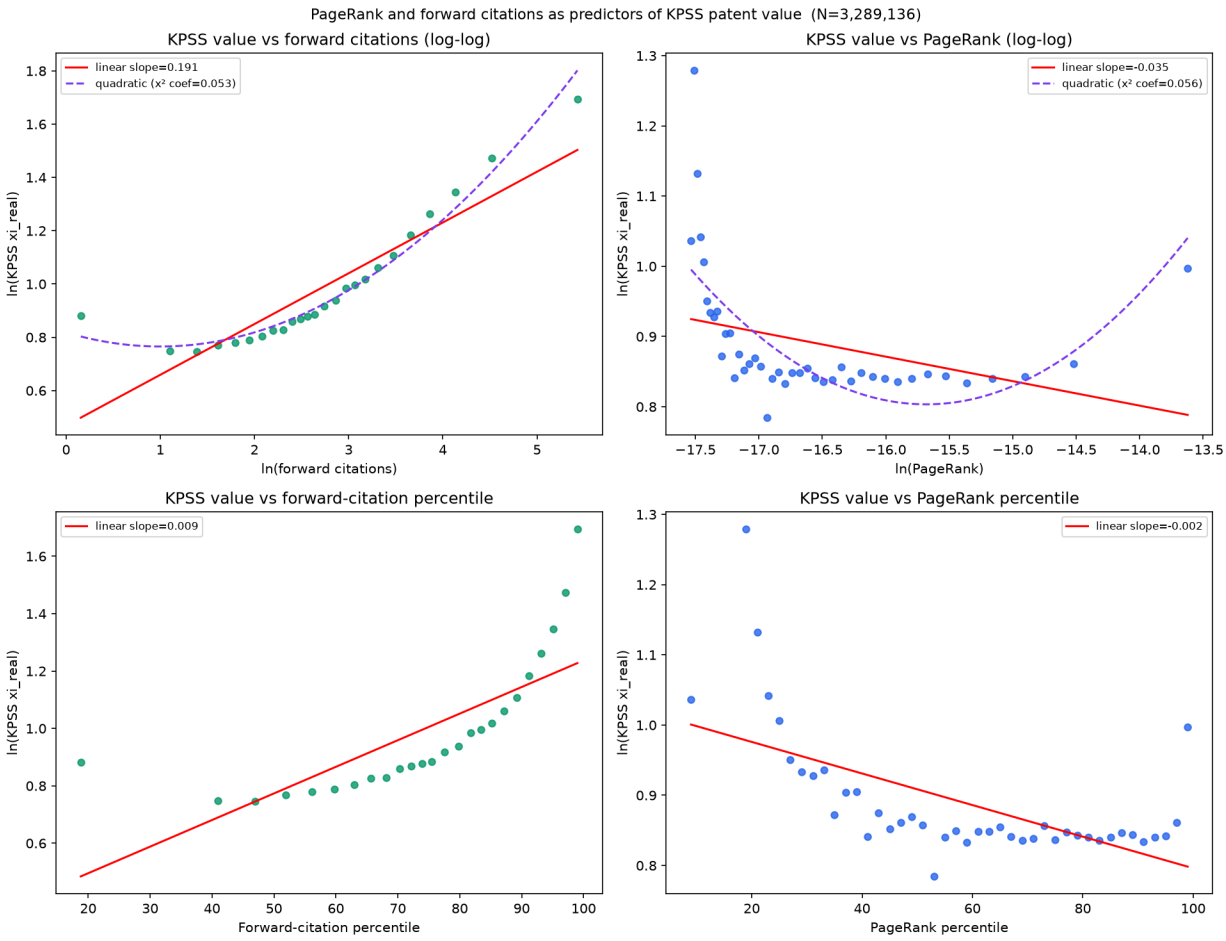


Figure A.17: KPSS Patent Value vs. Forward Citations and PageRank

Notes: Binned scatters (50 equal-count bins) of $\ln(\xi_{\text{real}})$, the KPSS stock-market patent value, against forward citations and PageRank. Top row: log-log with linear and quadratic fits. Bottom row: percentile ranks. Full KPSS/KPST regression tables are in this appendix.

Table A.9: OLS Regressions: KPST Text Importance on PageRank and Citations

	(1) (cit.)	(2) (raw PR)	(3) (raw PR + cit.)	(4) (age-adj. PR)	(5) (Mariani R_p)
ln(citations)	+0.0553	—	+0.0364	−0.0026	+0.0178
	(0.0001)		(0.0001)	(0.0001)	(0.0001)
ln(PageRank) (raw)	—	+0.0816	+0.0422	—	—
		(0.0001)	(0.0002)		
PageRank (age-adj., $\tilde{r}$)	—	—	—	+0.0486	—
				(0.0001)	
PageRank (Mariani R_p)	—	—	—	—	+0.0291
					(0.0003)
Filing-year FE	No	No	No	Yes	Yes
CPC-section FE	No	No	No	Yes	Yes
N	6,769,739	6,769,739	6,769,739	6,769,739	6,769,739
R^2	0.088	0.080	0.099	0.525	0.521

Notes: Dependent variable: KPST $\ell q_{sim,010}$ (Kelly et al., 2021). "Age-adj. PR" is the within-cohort standardized log-PageRank. Columns (4)–(5) use two-way Frisch–Waugh–Lovell demeaning; total R^2 reported. HC1 robust SEs.

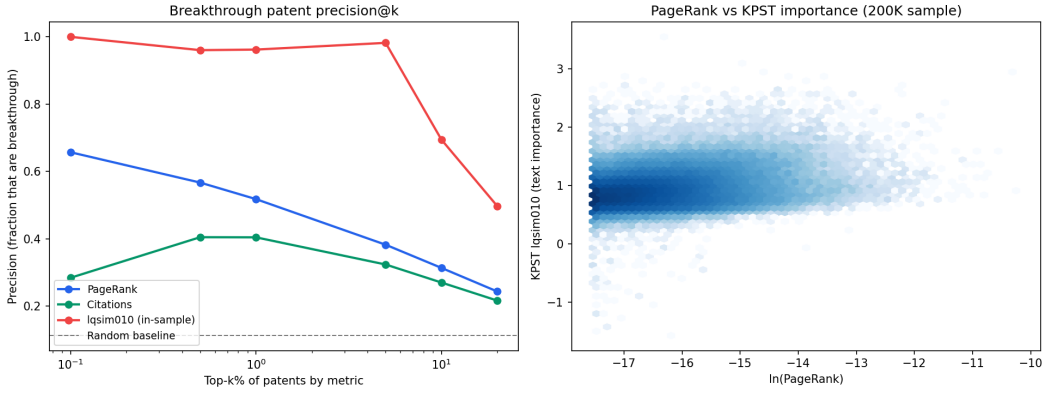


Figure A.18: Precision at k for KPST Breakthrough Patents: PageRank vs. Citations

Notes: Precision at k for KPST breakthrough patents (top decile of $\ell q_{sim,010}$). PageRank outperforms citation count at all thresholds; the gap is largest in the top 0.1–1%.

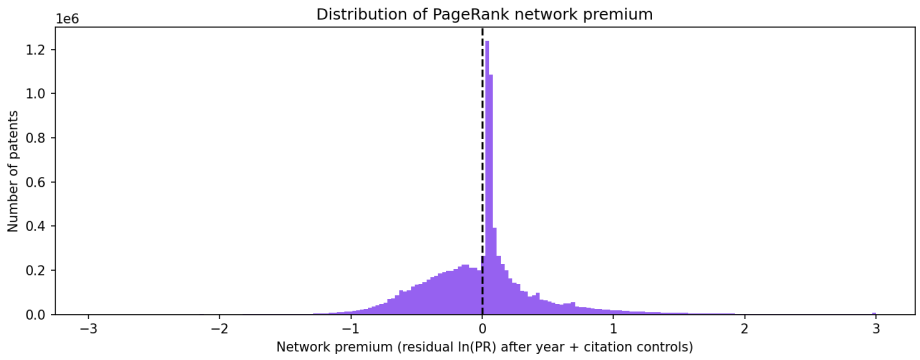


Figure A.19: Distribution of the PageRank Network Premium

Notes: Residual $\pi(i)$ from regressing ln(PageRank) on ln(citations) within filing-year cohorts. Positive: PageRank higher than citations/vintage predict.

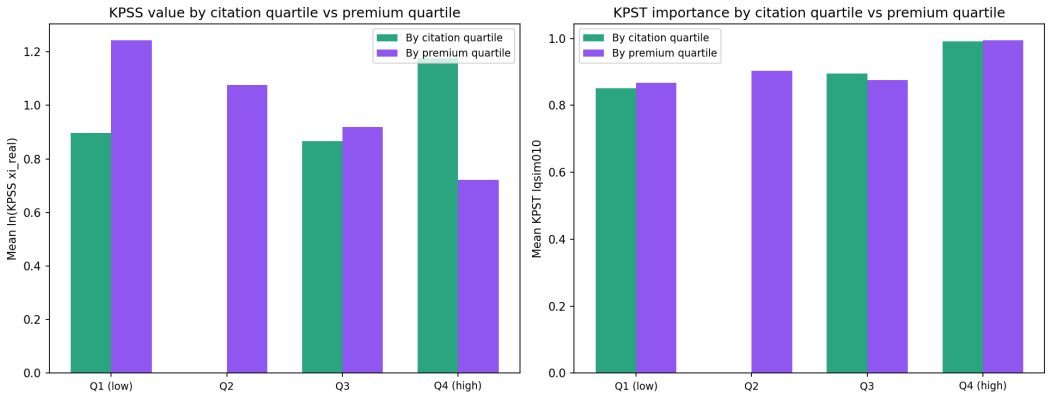


Figure A.20: Network Premium and External Quality Measures

Notes: Mean external quality (KPSS left, KPST right) by citation quartile vs. network-premium quartile. The premium gradient is steeper for KPST.

B.7 Data-Quality Checks

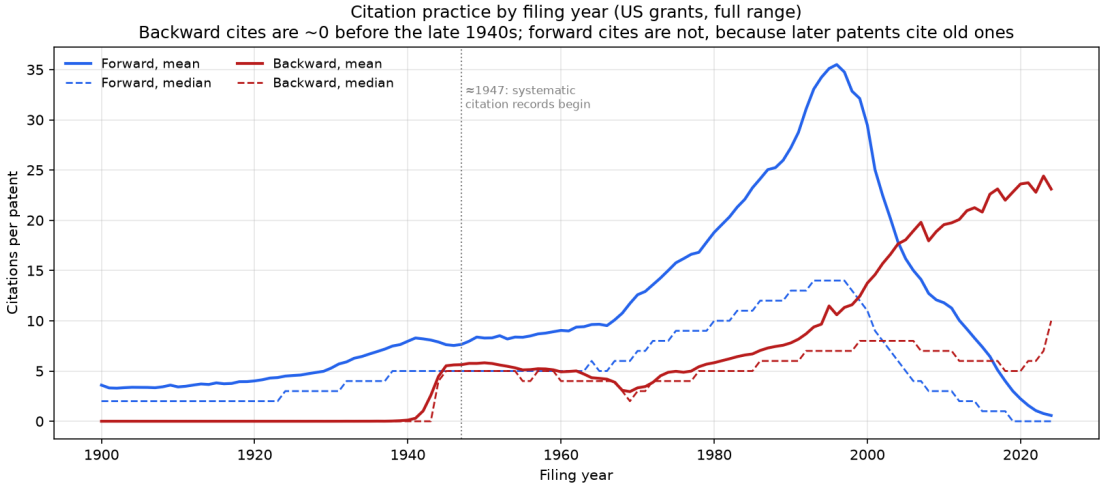


Figure A.21: Citation Practice by Filing Year (Data-Quality Check)

Notes: Forward and backward citations per US grant by filing year, full range. Backward citations are ≈ 0 before the late 1940s, consistent with systematic citation records beginning then; early forward-citation values are not a data error (later patents cite old ones).

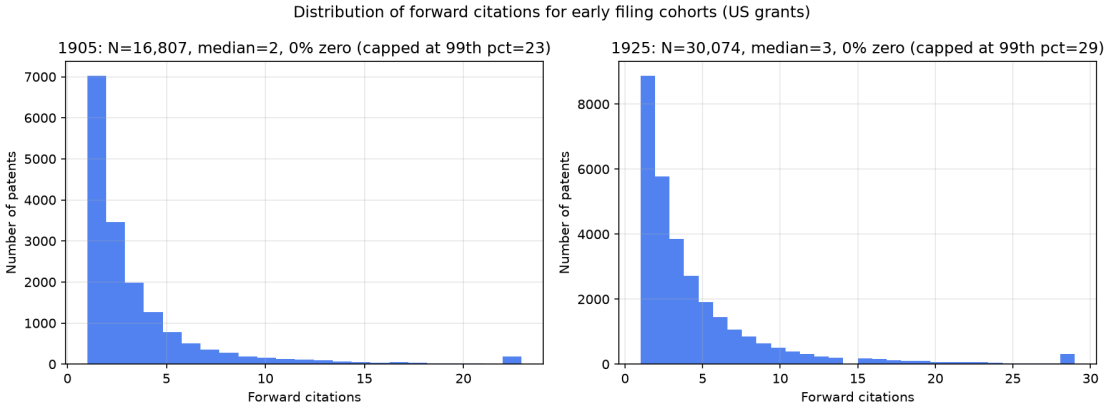


Figure A.22: Forward-Citation Distributions for Early Filing Cohorts (1905, 1925)

Notes: Forward-citation distributions for 1905 and 1925 cohorts (counts capped at the 99th percentile; zeros included).