The Dual Frontier:
Patented Inventions and Prior Scientific Advance

Mohammad Ahmadpoor[1,2] and Benjamin F. Jones[1,2,3]

[1] Northwestern University
[2] Northwestern University Institute on Complex Systems and Data Science
[3] NBER

May 27, 2017

Abstract

The extent to which scientific advances support marketplace inventions is largely unknown. We study 4.8 million U.S. patents and 32 million research articles to determine the minimum citation distance between patented inventions and prior scientific advances. We find that most cited research articles (80%) link forward to a future patent. Similarly, most patents (61%) link backwards to a prior research article. Linked papers and patents typically stand 2-4 degrees distant from the other domain. Yet advances directly along the patent-paper boundary are strikingly more impactful within their own domains. The distance metric further provides a typology of the fields, institutions, and individuals involved in science-to-technology linkages. Overall, the findings are consistent with theories that emphasize substantial and fruitful connections between patenting and prior scientific inquiry.

Scientific research can propel both fundamental understanding and practical application, but the extent to which scientific advances support technological progress is unclear (*1-3*). According to the "linear model" of science, basic research, focused on understanding, provides a foundation for eventual technological applications (*1, 4-7*). For example, Riemannian geometry, an abstract mathematical advance that was initially widely ignored, later proved essential to Einstein's development of general relativity and, ultimately, to time dilation corrections in the Global Positioning System. In biology, basic research into extremophile bacteria later proved essential to the development of the polymerase chain reaction, the DNA amplification technique that is vital to modern biotechnology applications. Such examples illustrate the potential value of the linear model as a conception of scientific and technological progress, a view that helps motivate the public case for supporting scientific research (*1, 8-9*).

At the same time, many observers argue that basic research rarely pays off in practical application or that practical advances typically proceed without any inspiration from basic research (*10-14*). These views suggest a potentially substantial disconnect between the knowledge outputs of public science institutions, such as research universities or government laboratories, and inventive outputs in the private sector. Other scholars argue for a richer interplay between scientific and technological progress. Characterizing scientific progress as advances in understanding and technological progress as advances in use, a common theme emphasizes that investigators focused on questions of use, engaged in solving real problems, may in turn generate new understandings and progress in basic science (*2, 15-17*). For example, Pasteur's germ theory of disease was closely intertwined with his work on industrial fermentation and food safety applications, and the development of the second law of thermodynamics was inspired by Carnot's practical interest in the efficiency limits of steam engines (*2, 7*). In these cases, new understandings of nature are seen less as independent exercises of human curiosity that pay off in unexpected, future applications, but rather as insights that spring up along the technological frontier.

Amidst these diverse views of the interplay between scientific and technological progress, there are many anecdotes but little systematic evidence. Our starting point is an integrated citation network that traces references from all 4.8 million patents issued by the U.S. Patent and Trademark Office (USPTO) from 1976-2015 to all 32 million journal articles published from 1945-2013 as indexed by the Web of Science (WOS), the world's largest collection of scientific research. The citation network begins by locating patents that directly cite journal articles, which defines a "paper-patent boundary" where practical inventions and scientific advances are linked (*18-21*). The network further determines the minimum citation distance for all other papers and patents to this boundary, creating a measure of distance that can be applied across a broad landscape of scientific and technological progress. We further integrate information about fields, individuals, and institutions (universities, government labs, and publicly-traded firms) for each paper and patent. The Supplementary Material (SM) details the underlying data sources and further discusses the use of citation networks to measure knowledge flows, including patent-to-paper citations (*22-26*).

Fig. 1A presents a schematic of the integrated citation network and introduces our metric. Formally, we define the distance metric $D_i \in \{1,2,3,\ldots\}$ for each patent or paper i. When a patent directly cites a paper, both nodes receive $D_i = 1$, representing patents and papers at the "patent-paper boundary". For the set of all other paper and patents, we recursively determine the minimum citation distance to this boundary. Namely, a paper i with $D_i = n + 1$ is one that is cited by a paper j with $D_j = n$ and is not cited by any paper k with $D_k < n$. Similarly, a patent i with $D_i = n + 1$ is one that cites a patent j with $D_j = n$ and does not cite any patent k with $D_k < n$. Paper and patents that cannot be connected at any distance to the paper-patent boundary are described as "unconnected". Note that the graph is directed: we trace citations backwards in time, using the references in each patent and paper and jumping from the patent to the paper domain where $D_i = 1$.

Our first results concern connectivity, considering the extent to which papers or patents exist in independent spheres. As shown in Fig. 1B, the patent-paper citation network has been dominated by a single connected component. A majority of patents – 60.5% – made references that could ultimately be traced to scientific papers. Similarly, among all scientific and engineering papers that received at least one citation, 79.7% could ultimately be connected to a patent. In short, we find *majority connectivity*, where the substantial majority of cited research articles can be linked to a future patent, and the modest majority of patents can be linked to prior scientific research.

At the boundary, 0.759 million patents directly cited 1.41 million papers, representing 21% of all connected patents and 10% of all connected papers (Fig. 1C). While these numbers are substantial, the broader picture that emerges in Fig. 1C is one of *indirect connectivity*. The modal connected science and engineering paper was 3 degrees from the nearest patent. The modal connected patent was 2 degrees from the nearest paper. Looking between 2 and 4 degrees of the patent-paper boundary captures 68% of all connected patents and 79% of all connected papers.

Our second set of results applies the distance metric to characterize fields. We used 185 WOS field classifications for science and engineering papers and the 388 primary USPTO technology classes that contained at least 20 patents each. For each field or class, Fig. 2A presents the mean distance, D_{mean}, among connected papers or patents as well as the percentage connectivity (i.e., the percentage of papers or patents in that field for which D exists). Here we see the enormous variation across fields. D_{mean} ranged from 2.00 to 5.90 across science fields and from 1.17 to 5.65 across patent classes.

Examining patents in Fig. 2A, the closest technology classes to the paper-patent boundary include combinatorial chemistry, molecular biology, superconducting technology, and artificial intelligence, all of which had $D_{mean} < 1.50$. The most distant technology classes concern subjects such as locks, buttons, fasteners, envelopes, fire escapes, and chairs, all of which had $D_{mean} > 4.75$. To further characterize this variation, we examined the full D distributions for several major technology classes (Fig. 2B). For example, we see that $D_{mode} = 1$ for "multicellular living organism" patents, where 85% directly cited papers, while $D_{mode} = 5$ for "chairs and seats" patents, for which only 0.3% directly cited papers.

Examining papers in Fig. 2A, we see that mathematics proved the most distant field from the patent frontier ($D_{mean} = 4.97$). Meanwhile, the closest fields to the patent frontier include nanotechnology, materials science & biomaterials, and computer science hardware & architecture, all with $D_{mean} < 2.35$. Fig. 2B provides the full D distributions for several major fields. Connected papers in mathematics, often considered a basic field of inquiry but one that can also be applied, had $D_{mode} = 4$ but with high variance. Astronomy and astrophysics also had $D_{mode} = 4$ but with a sharper peak and typically greater proximity to the patent-paper boundary. By contrast, biochemistry & molecular biology papers had $D_{mode} = 2$, and computer science papers had $D_{mode} = 1$, where 42% of connected computer science papers were directly cited by patents. This application to scientific fields suggests the potential usefulness of the distance metric for quantifying and tightening traditional but loose descriptors around "basic" and "applied" scientific research. The SM shows that the field ordering by distance to the patent-paper boundary is robust to different referencing tendencies across fields, to dropping patent-examiner citations in patents, and considers a null model (Figs. S1, S8, S9). Tables S1 and S2 provide the mean, mode, and standard deviation of the distance metric and percentage connectivity for all patent technology classes and all WOS fields.

Fig. S2 considers a related concept of distance: time. We calculated the total time period, T_i, in years along the shortest citation path between a paper and a patent. This time period is the difference between the patent's application year and the paper's publication year. At the boundary, where $D = 1$, there was a mean delay of 6.66 years. By $D = 6$, the mean delay was 19.62 years for papers and 22.70 years for patents. Fig. S2 further shows that the temporal distance varied substantially across fields, commensurate with the citation distance variation in Fig. 2A.

Fig. 3 considers impact. A common measure of impact for a scientific paper or patent is the number of citations it receives, and a transparent, field-independent metric considers the probability of a "home run," defined as being in the upper 5% of citations received in that field and year (*27-29*). Fig. 3A examines the probability of such home-run papers and patents. Patents that drew directly on scientific papers (i.e., $D = 1$ patents) were found to be unusually heavily cited *by other patents*, appearing as home runs 7.62% of the time, or 52.4 percent more often than the background rate. Other connected patents (i.e., $D \geq 2$ patents) were home runs at approximately the background rate. Fig. S3 shows more generally that impact decayed smoothly with distance from the frontier. Meanwhile, patents whose cited prior art was disconnected from the corpus of papers were home runs at a rate of 3.74%, or 25.2 percent less often than the background rate. Looking at papers in Fig. 3A, journal articles directly cited by a patent (i.e., $D = 1$ papers) were 3.72 times more likely to be highly cited *by other papers*. In other words, the patent-paper boundary appears populated by advances that are especially impactful within their own domains: patents that reference scientific papers were drawn on especially heavily by future patents, and papers cited directly by patented inventions were especially highly cited by other scientific papers. Meanwhile, patents or papers that were disconnected from the other knowledge network were especially unlikely to be high impact within their own domains.

The impact advantages are robust to numerous controls, including fixed effects for each year, field, number of authors (paper) or inventors (patent), institution type, and each number of references made by the paper or patent (Fig. S4). Fixed effect regressions account in a flexible and non-parametric manner for these features (see Methods in SM). Tables S3-S4 present the underling regression results and also show that these results are robust to alternative measures of citation impact. We also find similar results using patent maintenance fee payments rather than citations received (Table S5). Maintenance fees, which are paid by the patent owner and prevent the patent from lapsing, provide a potentially more direct measure of market value (*30-31*). Fig. S5 further shows that $D = 1$ patents didn't simply cite established, popular papers; rather, papers cited by a patent in the year the paper was published tended to become home runs within science over the ensuing years. We also find that $D = 1$ patents and papers were also far more likely to be home runs when looking within the outputs of a given inventor or author (Tables S6, S7). Examining individual fields, Fig. 3B shows that $D = 1$ patents and papers were the most highly cited within their own domains for the majority of scientific areas and technology classes. In science, 99% of fields, and in patenting, 86% of fields, showed that the highest impact work within the field occurs at $D = 1$.

Finally, we investigate the roles of institutions and individuals near the patent-paper boundary. Fig. 4A considers institutions. For comparison, we sorted relevant USPTO patents and WOS papers into three different institutional settings: universities, U.S. government laboratories, and firms. Institutional affiliations are based on patent assignee for the patents and based on postal and email addresses of the journal article authors (*32*-*33*). The SM provides additional details of this sorting process. Universities and government laboratories were relatively more engaged in high-D research whereas the research articles produced in firms shift towards $D = 1$ (Fig. 4A). These findings are consistent with and can help quantify long-standing ideas about the research outputs that for-profit institutions are likely to undertake (*34*). Table S8 provides associated regression analysis, including fixed effects for the number of references made, citations received, field, year, and number of authors or inventors. The regressions show that university papers were on average $D = 0.358$ further from the frontier than the firm papers. Decomposing this increased distance among university papers shows that approximately one-third of this increased distance was due to field composition (e.g., university researchers publish more in high-D fields like mathematics than corporate researchers do) and two-thirds appeared as institutional differences within a given field (e.g., university papers in mathematics have higher D than firms' papers in mathematics).

Fully 57% of university-assigned patents had $D = 1$, indicating the intensiveness of university patenting near the boundary (Fig. 4A). Patents from firms peaked at $D = 2$, with only 19% at $D = 1$. Patents by government laboratories appeared in between the other institutions. Table S9 provides associated regression analysis, showing that, compared to firms, approximately one-half of university patents' increased proximity to science was due to field composition (university researchers patented in low-D technology classes) and one-half appeared as institutional differences within a given field (e.g., university patents in material science had lower D than firms' patents in material science).

We next considered the institutional "hand-off" across the boundary where $D = 1$. For $D = 1$ patents, 78% were assigned to firms, yet 80% of $D = 1$ papers had university authors (Fig. 4B). The prevalence of hand-offs from university papers to business patents is consistent with long-standing conceptions that consider university outputs as public goods upon which marketplace invention can draw (*1*). Thus, while university patenting is particularly closely related to science (Fig. 4A) and can thus play a direct role in technology transfer (*35*-*36*), the lion's share of $D = 1$ patents still comes from firms. Related, other patents typically connected to the patent-paper frontier through these $D = 1$ firm patents (Fig. S6).

Fig. 4C examines the role of the same individual in spanning the paper-patent boundary. We define these cases by matching the inventor names for the patent with the author names for the paper that the patent cites (see SM for further discussion). For $D = 1$ university patents, 55.4% cited a paper written by an individual with the same name. A high percentage also appeared for government patents, but the percentage fell to 14.3% for $D = 1$ corporate patents. In Stokes' theoretical characterization of "Pasteur's Quadrant" (*2*), where the same individual may be engaged in advancing both understanding and use, universities and government labs appear to be especially common homes for such individuals, who in turn appear highly productive. Fig. S7 and Table S10 show that both the paper and the patent produced by such an individual were especially likely to be home runs in their respective domains.

Conclusions

Contrary to conceptions in which technological and scientific progress operate in independent spheres, we find majority connectivity between the corpus of patented inventions and the corpus of scientific papers. However, these connections are typically indirect, and both scientific fields and patenting technology classes vary enormously in their connectivity and proximity to the other domain. These findings are consistent with and can help quantify some features of the "linear model" of science, which imagines that scientists typically work to advance understanding but that such advances may underlie practical applications, often in indirect or unexpected ways. The prevalence of private-sector patents linking back to the output of universities and government laboratories is further consistent with institutional views of the linear model. While these features of the linear model appear to receive strong support, note that our data do not address potentially "non-linear" reverse linkages where technological advances, including new equipment and tools, may also drive scientific progress (*7*, *11*, *17*).

The distance metric further reveals facts that are consistent with and help quantify the fruitful, creative interplay between understanding and application (*2*, *19*, *21*). Patented inventions that draw directly on scientific advances were especially impactful compared to other patents. Moreover, papers directly cited by patents were also the highest impact papers within the scientific domain. These facts are consistent with a sharp complementarity between understanding and use and are also reflected at the individual level; an individual scientist/inventor, especially in university and government laboratory settings, often personally spanned the boundary, working to advance both the scientific and technological frontiers and managing to hit "home runs" in both domains.

Beyond loose classifications of "basic" or "applied" research and related terminologies (*6*, *7*), the distance metric provides a quantifiable typology to describe R&D outputs and the nature of their impacts. The typology can characterize the research outputs of not only fields, but also journals, funders, research institutions, and individuals themselves. Indices based on the D metric may thus present useful tools for understanding and evaluating types of research, institutional priorities, funding outcomes, and individual careers. While the distance metric in our application uses a directed graph, from patented invention to scientific advance, one may also deploy the metric on knowledge networks built using other link definitions. For example, full text analyses might allow one to characterize "necessary" precursor knowledge as opposed to the standard of "relevant" precursor knowledge that appear to be indicated by citation networks (see SM discussion). One might also build a metric that runs from scientific advances back to prior patented technologies, given appropriate reference information. And one might consider inventions or other applications outside patents. Such studies would further enrich our understanding of the interplay between scientific advance and technological progress to engage additional theories (*11*, *17*).

References

1. Bush, V., Science The Endless Frontier. A Report to the President. United States Government Printing Office, July, Washington, DC., (1945).
2. Stokes, D., Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Press, Washington, DC., (1997).
3. Lane, J., Bertuzzi, S., Measuring the returns of science investments. Science 331, 678-680, (2011).
4. Carty, J., The Relation of Pure Science to Industrial Research. Reprint and Circular Series, No.14, National Research Council, (1916).
5. Maclaurin, W.R., The sequence from invention to innovation and its relation to economic growth. Quarterly Journal of Economics 67, 97–111, (1953).
6. Godin, B., The linear model of innovation: the historical construction of an analytical framework. Science Technology & Human Values 31 (6), 639–667, (2006).
7. Balconi, M., S. Brusoni S., Orsenigo, L., In defence of the linear model: An essay. Research Policy 39, 1-13, (2010).
8. National Science Foundation, Basic Research: A National Resource, United States Government Printing Office, Washington, DC., (1957).

9. Nelson, R., The simple economics of basic scientific research, Journal of Political Economy 67, 297-306, (1959)
10. US Department of Defense, Project Hindsight Final Report. Office of the Director of Defense Research and Engineering, Washington, DC., (1969).
11. Kline, S.J., Rosenberg, N., An overview on innovation. In: Landau, R., Rosenberg, N. (Eds.), The Positive Sum Strategy. National Academy Press, Washington, DC., (1986).
12. Mansfield, E., Academic research and industrial innovation. Research Policy 20, 1–12, (1991).
13. Klevorick, A., Levin, R., Nelson, R., Winter, S., On the sources and significance of inter-industry differences in technological opportunities. Research Policy 24 (2), 185–205, (1995).
14. Von Hippel, E., The Sources of Innovation. Oxford University Press, Oxford, (1988).
15. Nelson, R., The link between science and invention: the case of the transistor. In: National Bureau of Economic Research (NBER), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press, Princeton, NJ, pp. 549–583, (1962).
16. Rosenberg, N., How exogenous is science? In: Rosenberg, N., Inside the Black Box: Technology and Economics. Cambridge University Press, Cambridge, 141–159, (1982).
17. Brooks, H., The relationship between science and technology. Research Policy 23, 477-486, (1994).
18. Narin, F., Hamilton, K. Olivastro, D., The increasing linkage between U.S. technology and public science. Research Policy 26, 317-330, (1997).
19. Fleming, L., Sorenson, O., Science as a map in technological search. Strategic Management Journal 25, 909-928, (2004).
20. Gaetani, R., Li Bergolis, M., The economic effects of scientific shocks. Northwestern University working paper, (2015).
21. Cassiman, B., Veugeleres, R., and Zuniga, P., In search of performance effects of (in)direct industry science links. Industrial and Corporate Change 17, 611-646, (2008).
22. Jaffe, A. B., Trajtenberg, B., Patents, citations, and innovations: A window on the knowledge economy, MIT Press, (2002).
23. Meyer M., Does science push technology? Patents citing scientific literature. Research Policy 29 (2000).
24. Tijssen, R. J. W., Buter, R. K., Van Leeuwen, Th. N., Technological relevance of science: An assement of citation linkage between patents and research papers, Scientometrics, Vol. 47, No. 2, (2000).
25. Hicks, D., Breitzman, T., Olivastro, D., Hamilton K., The changing composition of innovative acitivity in the US - a portrait based on patenet analysis, Research Policy, Vol. 30, Issue 4, (2001).
26. Callaert, J., Pellens, M. & Van Looy, B., Sources of inspirations ? Making sense of scientific references in patents. Scientometrics 98: 1617. (2014).
27. Wuchty, S., Jones, B., Uzzi, B., The increasing dominance of teams in production of knowledge. Science 316, 1036-1039, (2007).
28. Jones, B., Wuchty, S. Uzzi, B., Multi-university research teams: Shifting impact, geography, and stratification in science. Science 322, 1259-1262, (2008).
29. Wang, D., C. Song, Barabási, A. L., Quantifying long-term scientific impact. Science 342, 127-132, (2013).
30. Schankerman, M., Pakes, A., Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period. Economic Journal 96, 1052-1076, (1986).
31. Harhoff, D., Narin, F. Scherer, F. M., Vopel, K., Citation frequency and the value of patented inventions. Review of Economics and Statistics 81, 511-515, (1999).
32. Arora, A., Belenzon, S., Killing the golden goose? The decline of science in corporate R&D. NBER Working Paper No. 20902, (2015).
33. Bryan, K., Ozcan, Y., The impact of open access mandates on invention. University of Toronto Working Paper, (2016).

34. Rosenberg, N., Nelson, R., American universities and technical advance in industry. Research Policy 23, 323-348, (1994).
35. National Academy of Sciences, Managing university intellectual property in the public interest. National Academy Press: Washington, DC., (2010).
36. Thursby, J., Thursby, M., University licensing and the bayh-dole act. Science 301, 1052, (2003).
37. Jaffe, A., Trajtenberg, M. and R. Henderson, Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. Quarterly Journal of Economics 108(3), 577-598, (1994) .
38. Furman, J. and S. Stern, 2011. Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production. American Economic Review 101(5), 1933-1963.
39. Azoulay, P., Graff Zivin, J. and J. Wang, 2010. Superstar Extinction. Quarterly Journal of Economics 125(2), 549-589.
40. Uzzi, B., Stringer, M., Mukherjee, S., Jones, B., 2013. Atypical combinations and scientific impact. Science 342, 468-472.
41. Meyer, M. & Persson, O. Nanotechnology-interdisciplinarity, patterns of collaboration and differences in application Scientometrics (1998)
42. Acemoglu, D., U. Akcigit, and W. Kerr. 2016. Innovation Network. Proceedings of the National Academy of Sciences 113(41), 11483-11488.
43. Boyack, K., R. Klavans, and K. Borner. 2005. Mapping the Backbone of Science. Scientometrics 64(3), 351-374.
44. Almeida, P., 1996. Knowledge Sourcing by Foreign Multinationals: Patent Citation Analysis in the U.S. Semiconductor Industry. Strategic Management Journal 17(52), 155-165.
45. Breschi, S. and F. Lissoni. 2005. Cross-Firm Inventors and Social Networks: Localized Knowledge Spillovers Revisited. Annales d'Economie et de Statistique 79/80, 189-209.
46. Alcacer, J., Gittleman, M., Patent Citation and Measure of Knowledge Flows: The Influence of Examiner Citations, Review of Economics and Statistics (2006)
47. Aknes, D. 2005. Citations and Their Use As Indicators in Science Policy. Ph.D. Dissertation. Mimeo, University of Twente.
48. Jaffe. A., M. Trajtenberg, and M. Fogarty, Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. American Economic Review 90(2), 215-218, (2000).
49. Finardi, U., Time Relations between Scientific Production and Patent Knowledge: The Case of Nano-technologies, Scientometrics, Volume 89, Issue 1, (2011).
50. Lo, S., Scientific linkage of science research and technology development: A case of genetic engineering research, Scientometrics, Volume 82, Issue 1, (2010)
51. Nagaoka, S. and I. Yamauchi. 2015. The Use of Science for Inventions and its Identification: Patent Level Evidence Matched with Survey. RIETI Discussion Paper Number 15-E-105.
52. Lai, R., A. D'Amour, A. Yu, Y. Sun, and L. Fleming. 2011. Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010). Harvard Dataverse, V5, https://dataverse.harvard.edu/dataverse.xhtml?alias=patent.

Acknowledgements

We gratefully acknowledge support from the Alfred P. Sloan Foundation under award G-2015-14014. We thank workshop participants at National Academy of Sciences, the European Policy for Intellectual Property Association, the Northwestern Institute for Complex Research, the Institute for Policy Research, Adam Jaffe, Alan Marco, Manuel Trajtenberg, and Brian Uzzi for helpful comments. We are especially grateful to Ruben Gaetani for help with patent data. The Web of Science data is available via Thomson Reuters. The patent datasets are available publicly as discussed in SM. The constructed D metric variables are available from the authors. Contact: bjones@kellogg.northwestern.edu (corresponding author) and mohammad.ahmadpoor@kellogg.northwestern.edu.

Supporting Online Material

Materials and Methods
Figs. S1 to S9
Tables S1 to S10
References (39-5)

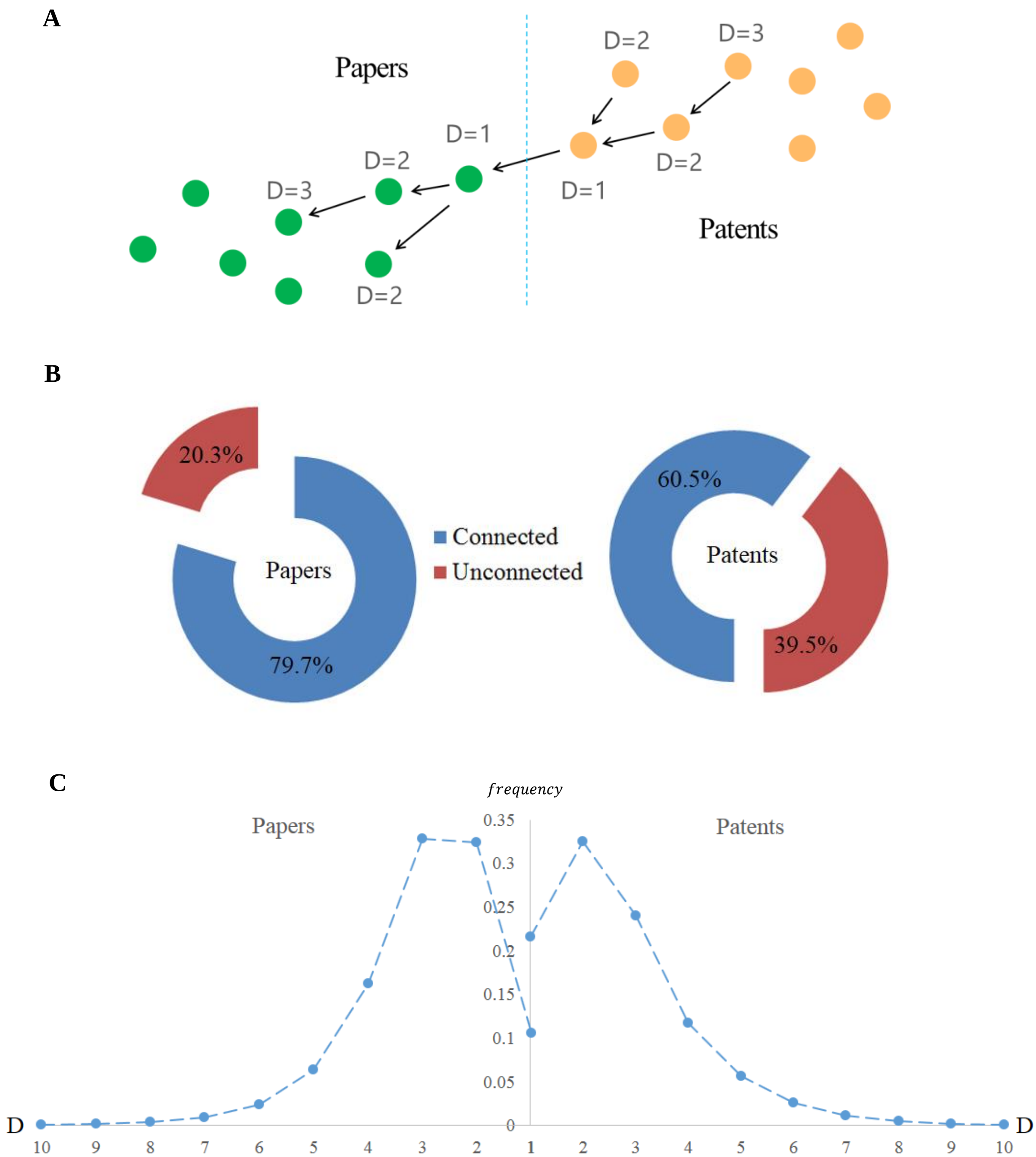


Fig. 1. Connectivity and distance. (**A**) The directed graph of the integrated citation network from patents toward papers defines a distance metric, D. (**B**) The share of papers that link forward to a future patent and the share of patents that link backward to a prior research article. (**C**) The distance distribution of connectivity.

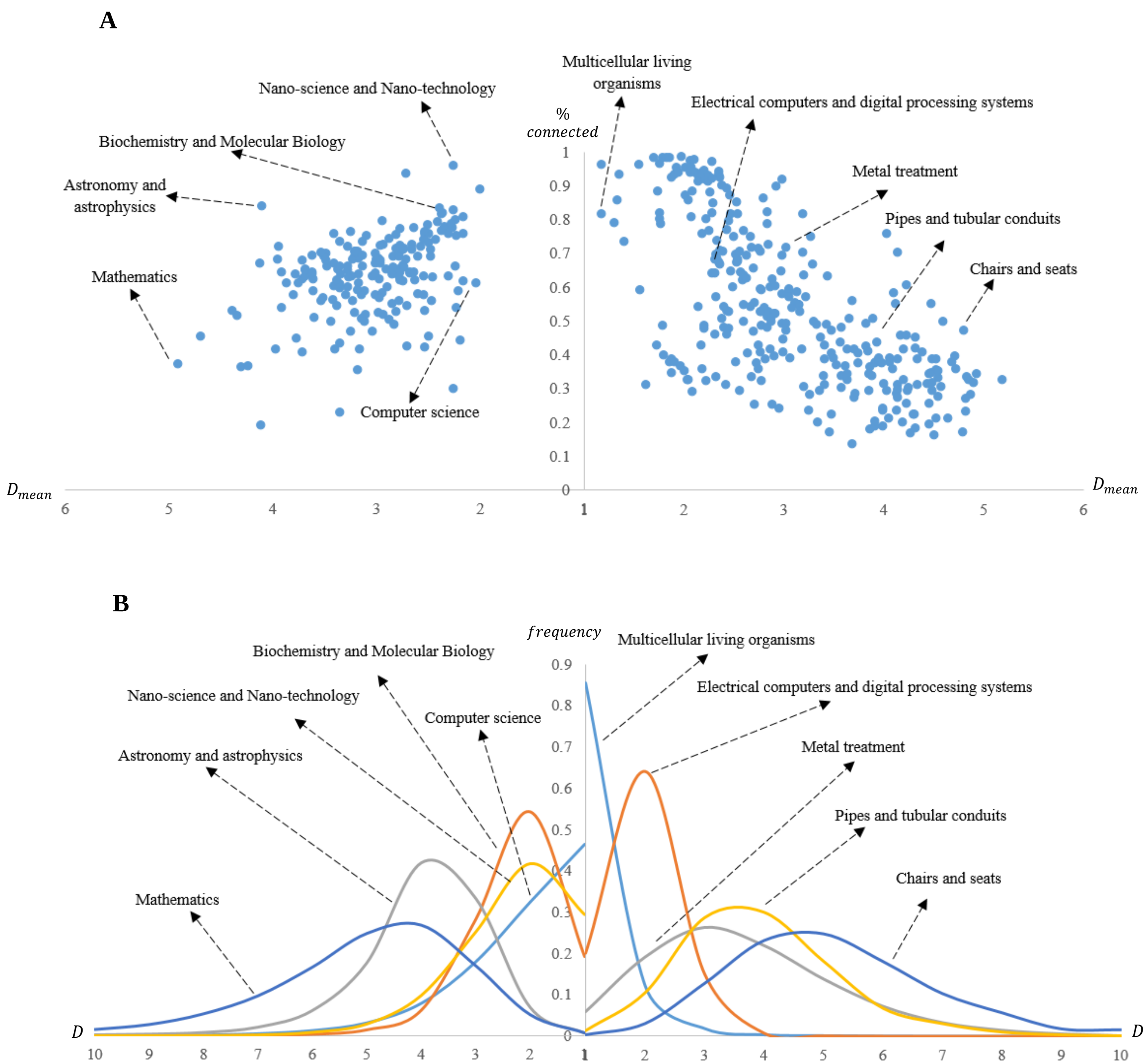
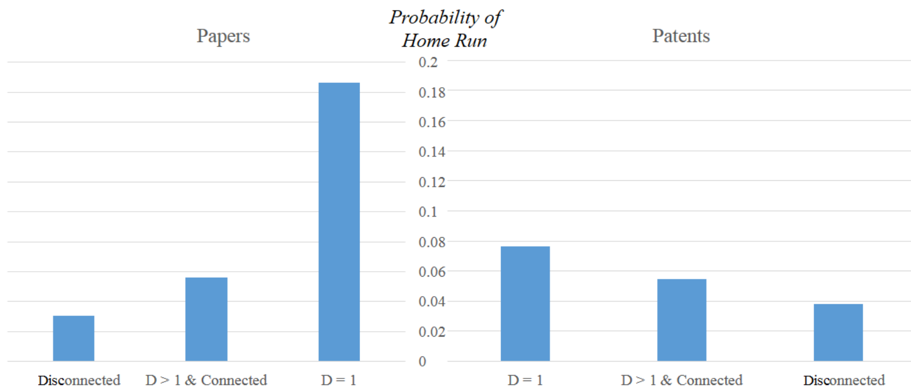


Fig. 2. Application to Fields. (**A**) Distance metric. The mean distance, D_{mean}, to the paper-patent boundary is presented for each field (x-axis) together with the percentage of knowledge outputs in that field that are connected to the integrated citation network (y-axis). (**B**) The full D distribution for several fields.

A



B

	Distance where Home Run Rate is Largest, by Field or Class		
	$D = 1$	$D > 1$	Disconnected
Papers (WOS fields)	99%	1%	0%
Patents (USPTO classes)	86%	14%	0%

Fig. 3. Distance and Impact. (**A**) Impact close to and far from the paper-patent boundary. A "home run," is defined as being in the upper 5% of citations received in that field and year, for a patent or a research paper. (**B**) Home runs outcomes relative to distance for each field, when each field is analyzed separately. The SM examines alternative impact measures, including methods based on patent renewal payments.

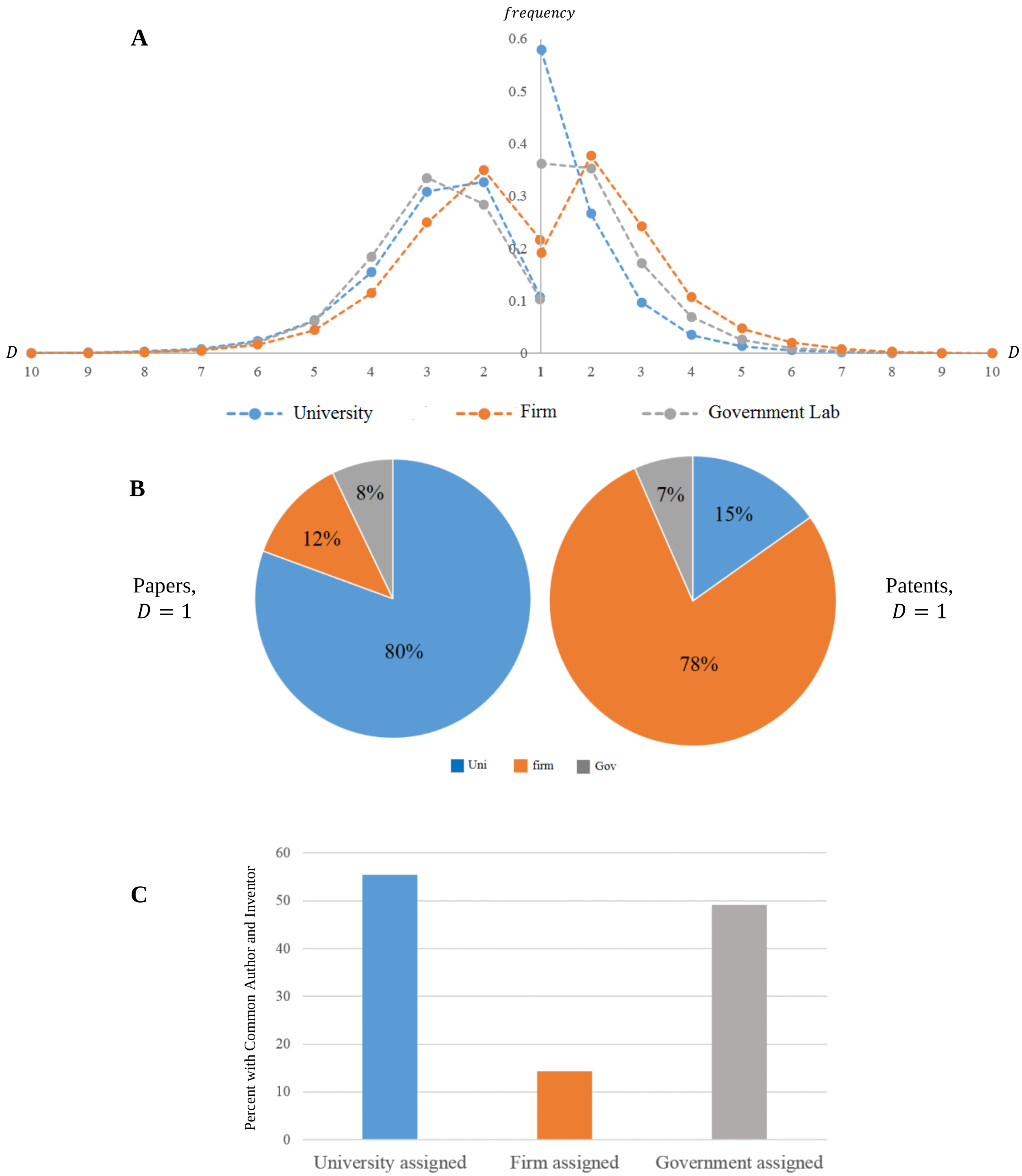


Fig. 4. Institutions and Individuals. (**A**) The D distribution for different institutional settings, including universities, government laboratories, and firms. (**B**) Production of patents and papers by institutional type at the $D = 1$ boundary. (**C**) The share of $D = 1$ patents where a citing inventor and cited author have the same name, by patent assignee type.

Supplementary Materials for

The Dual Frontier:
Patentable Inventions and Prior Scientific Advance

Mohammad Ahmadpoor[1,2] and Benjamin F. Jones[1,2,3]
[1] Northwestern University
[2] Northwestern University Institute on Complex Systems and Data Science
[3] NBER

Correspondence to: bjones@kellogg.northwestern.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S9
Tables S1 to S10

Materials and Methods

DATA

Our measurement approach builds on large literatures that use patent citations and paper citations to trace knowledge flows (e.g., *37-40*) and literature examining references in patents to publications (e.g., *18-19*, *41*). The integrated citation network introduced in this paper merges patent and paper datasets.

Patent Data

We studied all 4.8 million patents granted by the United States Patent and Trademark Office (USPTO) between 1976 and 2014. These data are drawn from overlapping datasets, including the Patent Data Project (https://sites.google.com/site/patentdataproject/Home) of the National Bureau of Economic Research and the updated patent data of Kogan et al. (2015) (https://iu.app.box.com/v/patents).

Together, these data record the patent number, application year, patent references, inventor names, assignee (owner), and the technological class of the patent. In all of our analyses, we use the application year to locate the patent in time. Technological class for each patent is determined by the USPTO, which identifies 430 different primary classes. Our analysis focuses on the 388 classes that have more than 20 patents in our citation network.[1]

Patent renewal data is obtained from a USPTO database that records maintenance fee events (which occur in the 4th, 8th, and 12th year after the patent was granted). This data is available for patents granted from September 1, 1981 to the present and is available here: https://bulkdata.uspto.gov/data2/patent/maintenancefee/

Paper Data

We examined 32.4 million scientific publications, constituting all research articles indexed in the Thomson Reuters Web of Science (WOS) database that were published over the 1945-2013 period. The WOS records paper titles, bibliographic information (journal, volume, issue, page), citations, author information (names, affiliations), and citation links to other papers in the database. From the total of 32.4 million journal articles, only the publications that are cited at least once can fall into our citation network (17.0 million papers). We build the integrated networks with this full set of data, which includes science and engineering, social science, and arts and humanities fields, and then focus the analysis on the 23.7 million journal articles in science and engineering fields, as codified by the WOS. These science and engineering papers are categorized by the WOS

[1] In Fig. 2C, we further focus on the 306 classes that have more than 10 patents at D=1.

into 185 different subfields.[2] The WOS data are available to researchers through Thomson Reuters and described in detail at www.webofknowledge.com.

Patent-Paper Boundary

Patent citations to WOS articles were provided by Gaetani and Li Bergolis (*20*), who used a full-text patent XML database from the USPTO to match non-patent references in the patents to WOS articles.[3] Matching was based on name of first author, journal, publication year, article title, volume, and page numbers.

In the WOS, the bibliographic information is organized in the database into name, journal, title, etc. In the USPTO reference, the reference information is contained in a single string. The matching algorithm of Gaetani and Li Bergolis (*20*) thus, as a first step, extracts the name of the first author and publication year from the patent reference and then locates the subset of WOS papers that match this information. In the second matching step, the WOS paper is selected that has the closest match to the USPTO reference string based on volume and page numbers and shared words in the journal name and paper title.

In some robustness tests, we will also consider the patent citation network dropping references in patents that were added by patent examiners. Patent-examiner added citations are denoted in the XML files and easily identified for patents issued in or after the year 2001.

METHODS

1. Citation linkages

The D-metric builds from the integrated citation network of U.S. patents and Web of Science journal articles. The analytic methodology for calculating the D-metric is provided in the main text. Here we provide further background, based in existing literature, regarding the uses and interpretations of citations linkages.

Citation Linkages and Knowledge Flows

Using citation linkages to inform knowledge flows is a core methodology in existing literature (*22*). Studies use citations to study knowledge flows and spillovers across space (e.g., *37*), over time (e.g., *42*), across fields (e.g., *43*), across organizations (e.g., *44*), and through social networks (e.g., *45*). Other work uses citation linkages to inform how prior knowledge is combined into new knowledge (e.g., *40*).

While the use of citations to study knowledge flows is common in the existing literature, it is important to recognize that citations, linking a new knowledge output to a specific, prior knowledge output, is a proxy measure and may have multiple

[2] Results for non-science and engineering fields are available from the authors upon request.
[3] Full text XML patent data are available from the USPTO at https://bulkdata.uspto.gov/.

interpretations. For example, in patents, citations may be used to delineate property rights vis-à-vis relevant prior work, which can be distinct from denoting important creative inputs (e.g., *46*). With papers, references may be added to inform referees (*47*). A related question is whether one should treat all references as equally important, even when they represent knowledge flows, as some prior references may be more consequential than others.

While citation linkages are the subject of ongoing research, we note a few additional points here relevant to our study. First, for patents, we consider several robustness analyses below restricting references in patents to those added by the applicant as opposed to the examiner, where the applicant-added references presumably come closer to demarcating knowledge flows from the inventor's perspective. Second, our analyses largely consider groups of knowledge outputs (by field, by institution, etc.), which may help avoid some problems with noise at the individual patent or citation level (*48*). Lastly, we emphasize that the D metric is flexible in that it can be applied to other definitions of knowledge links as these develop in future literature. For example, there are ongoing efforts to use full text analysis to determine key prior ideas within a given patent or paper. With the advent of such new link definitions, future research can deploy the D methodology upon them.

<u>Scientific Non-Patent References</u>

The study of non-patent references to scientific literature (sNPRs) has been previously conducted in studies of specific fields or smaller samples of patents, and with an emphasis on the immediately linked patent or paper. In the language of the D metric, an sNPR occurs at $D = 1$, as opposed to the full range of D studied in this paper. Narin et al. (*18*) study sNPRs in U.S. patents issued in 1987-88 and 1993-94. They find that the frequency of sNPRs is increasing and that the cited papers in their sample typically come from publically-supported research. Hicks et al. (*25*) study 50,000 sNPRs linking private-sector patents from 1993-1997 to non-private sector papers and show that these linkages are often geographically localized. Looking at specific fields, Finardi (*49*) studies the time lag between nanotechnology patents and the publication year of these patents' sNPRs, and Lo (*50*) shows that genetic engineering patents draw the majority of their prior art citations from papers rather than other patents. Fleming and Sorenson (*19*) study U.S. patents issued in May-June of 1990 and show that the patents with sNPRs are more highly cited than other patents. By contrast, Cassiman et al. (*21*) study approximately 1,000 EPO patents issued to 79 Flemish firms over 1995-2001 and argue that sNPRs do not predict greater future patent citations. Gaetani and Li Bergolis (*20*), who provide the sNPR data match we use between the full set of U.S. patents and the Web of Science, study patent references to approximately 200 extremely high impact research articles to examine the effect of scientific breakthroughs on the performance of the patenting firms.

Regarding the interpretation of sNPRs, several authors have studied the meaning of these linkages. Meyer (*23*), using a case study methodology of nanotechnology, suggests that sNPRs can represent a simultaneity of scientific and inventive output within the same individual, so that knowledge flows are moving in both directions (a theme we discuss vis-à-vis Stokes (*2*) in the paper). Meyer (*23*) also concludes that sNPRs are a general indicator of the science-relation of a field. Tjissen et al. (*24*) study patent citations to Dutch research papers and conclude that "these citations reflect genuine links between science and technology." Nagaoka and Yamauchi (*51*) conduct a survey of 843 inventors and find both that there are important linkages to science and that patent citations to scientific literature are an incomplete and noisy indicator of the knowledge flow, with errors of both over- and under-inclusion. Callaert et al. (*26*) interview 36 Belgian inventors in nanotechnology, biotechnology, and life sciences, and find among their EPO patents that only 20% of the sNPRs are described as "unimportant" by the inventor, whereas 34% are described as "background" while 44% are described as "important" or "very important". Hicks et al. (*25*) argue that the tendency for geographically localized linkages in their sample of sNPRs suggests that these are substantive spillovers. As we will show below, fully 96% of sNPRs after 2001 in USPTO patents are provided by the applicant, not the patent examiner, which may further suggest these are relatively substantive linkages. Overall, while this literature is still developing, sNPRs appear to a substantive if noisy indictor of the role of specific, prior scientific advances.

Patent Examiner Added Citations

Patent references can be provided by both (a) the applicant and (b) the examiner (*46*). In both cases, the cited references indicate some type of relevant prior knowledge. However, the two different sources of citation may suggest different interpretations. For example, patent-examiner added references are likely to be less important for understanding the material used in the inventor's creative process.[4] For patents issued in the year 2001 or later, the XML patent database (see above) directly identifies which references were added by examiners, both for the patent references and the non-patent references in each patent document. To help narrow interpretations and perform robustness checks, we have therefore further explored the citation network when patent-examiner added citations are eliminated.

We find that 36% of patent-to-patent references are added by examiners. By contrast, only 4% of patent-to-publication references are added by examiners. Thus an immediate observation is that, when patent examiners add references, they are far more likely to add a reference to other patents as prior art and rarely add references to other publications. For the network, it follows that (a) the identity of $D = 1$ patents and papers is driven almost entirely by applicant-added references; (b) the D metric is little changed for papers in general when dropping patent-examiner citations, because the identity of $D = 1$ papers changes little and papers with $D > 1$ are determined through paper-to-paper references; and (c) the D metric may shift more for patents with $D > 1$, as more of the patent-to-patent references are (36%) added by examiners.

Given the proportion of patent-to-patent references added by examiners, we further consider how the D distribution for patents appears when the patent-examiner added

[4] Although there may be under-inclusion of prior references by applicants and it is thus possible that the inventor may have used the prior advance without citing it; see, e.g., Nagaoka and Yamauchi (2015).

citations are dropped. Dropping a reference from a patent's reference list can have two possible effects: (i) it can cause a patent to become disconnected or (ii) it can cause the D for that patent to either stay the same or rise. The findings are as follows. First, among post-2001 patents, 44.7% of the patents are disconnected using a network built only from applicant-added references, which is a similar proportion when studying the full sample average as in Fig. 1. Second, and conditional on being in the connected network, Fig. S9A compares the D distribution for patents issued after 2001, both when patent examiner references are included in building the network and when they are not. We see that peaking behavior flattens somewhat at $D = 2$ and $D = 3$ using only applicant-added references, but overall the shape of the distribution is broadly similar. Finally, Fig. S9B compares the D_{mean} for each technology class, both when patent examiner references are included in building the network and when they are not. We see that the field ordering in terms of distance is largely stable.

2. Institution and Individual Matching

Institutional definitions and matching

Three categories for institutions involved in patenting and publishing articles were considered: (a) universities, (b) government laboratories, and (c) firms.

a) Universities: To find articles that are published from a university, we used the method of Bryan and Ozcan (*33*). In particular, we explored the author(s) addresses provided by the WOS database and searched for one of the following strings in the address entry: university , alumni , univ , national cancer , brigham , jackson lab , research center , akademie , vib , RIKEN , Eye & Ear , medical school , national jewish health , eth zurich , Center for , univeristy , higher education , cold spring harbor , akadamie , centre for , fundacio , Université , centre , planck , universuty , Universitât , fundacion , UNIVERSITÀ , agence nationale , insitute , UNIVERSITÉ , eye and ear in rmary , Society for , Unversity , cancer centre , universite , institue , istituto , cancer center , fondation , universiteit , universitet , universitaet , city of hope , educational fund , zentrum , consejo , ecole , universtiy , centro , kettering , mayo , schule , institucio , centrum , hospital for sick , children's hospital , academisch , universita , universit 'at , unviersity , georgia tech , school of , consiglio nazionale , intellectual properties , fondazione , national centre , centro nacional , centre national , foundation , regents , council , fred hutchinson , general hospital corporation , universidade , research hospital , medical center , foundation , universitat , universidad , colegio , univerisite , institut , institute , instituto , trustees , academia , academy , or college. The same procedure was followed for patents and we looked for the aforementioned strings in the assignee entry of patents in the patent data.

b) Government laboratories: To find papers that are published in government labs, we used a list of government labs provided by the NSF and searched for these labs in the author(s) address entry. The list consists of the following government labs: Aerospace Federally Funded Research and Development Center , Ames Laboratory, Argonne National Laboratory, Arroyo Center, Brookhaven National Laboratory, Center for

Advanced Aviation System Development, Center for Communications and Computing, Center for Enterprise Modernization, Center for Naval Analyses, Center for Nuclear Waste Regulatory Analyses, CMS Alliance to Modernize Healthcare, Fermi National Accelerator Laboratory, Frederick National Laboratory for Cancer Research, Homeland Security Studies and Analysis Institute, Homeland Security Systems Engineering and Development Institute, Idaho National Laboratory, Jet Propulsion Laboratory, Judiciary Engineering and Modernization Center, Lawrence Berkeley National Laboratory, Lincoln Laboratory, Los Alamos National Laboratory, National Biodefense Analysis and Countermeasures Center, National Center for Atmospheric Research, National Cybersecurity Center of Excellence, National Optical Astronomy Observatory, National Defense Research Institute, National Optical Astronomy Observatory, National Radio Astronomy Observatory, National Renewable Energy Laboratory, National Security Engineering Center, National Solar Observatory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, Princeton Plasma Physics Laboratory, Project Air Force, Sandia National Laboratories, Savannah River National Laboratory, Science and Technology Policy Institute, SLAC National Accelerator Laboratory, Software Engineering Institute, Systems and Analyses Center, and Thomas Jefferson National Accelerator Facility. To identify patents that have a government lab as an assignee, we searched for the aforementioned strings in the assignee entry of patents in the patent data.

c) Firms: For papers, we looked for journal articles that have one of the following strings in their corresponding author(s) address: Inc, Group, Foundation, Co, limited, LTD, LLC, Corp, Company, LP, and LLP. For the patent side, we used the NBER PDP Project data which matches patent data to Compustat firms. Note that this patent dataset concludes in 2006, so the institutional analyses conclude in that year.

Individual name matching at $D = 1$

The inventor names for patents are obtained from the NBER Patent Data Project (see Data above). The author names for papers are obtained from the WOS. For a patent that directly cites a paper, we considered whether any inventor on the patent shares the same name as any author on the paper. Names are matched based on last name and first initial. The idea of this algorithm is that, while names themselves may be difficult to disambiguate across millions of patents and papers, it is very rare that a person with a given name directly cites a person with the same name that is not himself/herself (*27*).

Name Disambiguation

In regression analysis, we consider models with fixed effects for the specific inventor or author. These regressions ask whether a given individual's output is higher impact, compared to that same individual's other output, when the output is at the patent-paper boundary.

To run these regressions, we need individual identifiers for inventors and authors. Name disambiguation is a well-known challenge across many domains. For patents, we use the Lai et al. (*52*) name-disambiguated inventor database, which to our knowledge is the state of art among publicly available inventor data. For papers, we create individual identifiers based on the author name (last name and first initial) and WOS subfield.

3. Regression Methods

Regression Methods for Patent Impact

Regression analyses of high-impact patents employ fixed-effect ordinary least squares models. These take the following form.

$$y_i = \beta_x x_i + \beta_w w_i + \sum_r \beta_r R_{ri} + \sum_z \beta_z Z_{yi} + \sum_n \beta_n N_{ni} + \sum_f \beta_f F_{fi} + \sum_a \beta_a A_{ai} + \epsilon_i$$

where i indexes a specific patent. The variables are defined as follows.

Dependent variable for patents: The dependent variable measures impact. In the main analysis (Fig. 3), we define a binary variable $y_i \in \{0,1\}$, where $y_i = 1$ indicates that the patent is in the upper 5th percentile of citations received compared to other patents with the same technological class and application year, and $y_i = 0$ otherwise, where citations are counted within the first 8 years after the patent grant (*27-28, 40*). Note that the citation counts only include citations from within the domain of patents (patents citing patents). Similarly, for papers (see below), we count citations only from other papers to isolate impact within the paper domain. Alternative measures considered below include using (i) an alternative binary variable using the upper 1st percentile of citations received as the threshold to define high impact, (ii) the log of citations received within 8 years after application, and (iii) an integer count of the number of patent renewal fees paid for the patent (*30-31*).

Predictors of interest: We examine in regression the extent to which $D = 1$ patents predict high impact, defining a binary variable $x_i \in \{0,1\}$, where $x_i = 1$ if $D_i = 1$ and $x_i = 0$ otherwise. Similarly, we examine unconnected patents (for which D is not defined), defining a binary variable $w_i \in \{0,1\}$, where $w_i = 1$ if $D_i = missing$ and $w_i = 0$ otherwise.

Fixed effects: To control for other possible influencers of impact and distance in a flexible manner, we include detailed fixed effects to account as follows.

R_{ri}: These fixed effects account for the number of references the patent makes. In particular, the R_{ri} are a series of individual binary variables $R_{ri} \in \{0,1\}$, where $R_{ri} = 1$ if patent i makes exactly r references and $R_{ri} = 0$ otherwise. In practice, we use an individual fixed effect for each integer number of references up to 100 and then bin the few patents that make 100 or more references as one category.

Z_{zi}: These fixed effects account for the institutional setting for the patent, based on the patent assignee. In particular, the Z_{zi} are four individual binary variables $Z_{zi} \in \{0,1\}$

for four possible institutional categories, $z \in \{u, g, f, o\}$, where $Z_{zi} = 1$ if patent i has assignee type z, and $Z_{zi} = 0$ otherwise. The four assignee types are universities (u), government laboratories (g), publicly-traded U.S. firms (f), or other (o). The other category indicates that the algorithm described above (see "Institutional definitions and matching" under Methods) did not classify the patent to one of the other three categories.

N_{ni}: These fixed effects account for the number of inventors on the patent. In particular, the N_{ni} are a series of individual binary variables $N_{ni} \in \{0,1\}$, where $N_{ni} = 1$ if patent i has exactly n inventors and $N_{ni} = 0$ otherwise. We use an individual fixed effect for each integer number of inventors.

F_{fi}: These fixed effects account for the technological class of the patent. The F_{fi} are a series of individual binary variables $F_{fi} \in \{0,1\}$, where $F_{fi} = 1$ if patent i is in technological class f and $F_{fi} = 0$ otherwise.

A_{ai}: These fixed effects account for the application year of the patent. The A_{ai} are a series of individual binary variables $A_{ai} \in \{0,1\}$, where $A_{ai} = 1$ if patent i has application year a and $A_{ai} = 0$ otherwise.

Regression Methods for Paper Impact

Regression analyses of high-impact papers employ the same fixed-effect ordinary least squares model defined for patents above. The variables are also adjusted for the different data. Now impact (the dependent variable, y_i) refers to citations received *by other papers*. The predictor variables are whether the paper is cited directly by a patent (a $D = 1$ paper) and whether the paper is disconnected from the patent-paper citation network. For the fixed effects, the number of inventors is replaced by the number of authors in the N_{ni}, the technological class is replaced by the WOS field code in the F_{fi}, and the application year is replaced by the paper's publication year in the A_{ai}.

Regression Methods with Individual Fixed Effects

Regression analyses of high-impact patents alternatively employ fixed-effects for each individual inventor or author. The Data section above describes the definition of the individual indicators. The data in these regressions takes the set of connected patents (or papers) and considers patent lists for each inventor (or paper lists for each author).

The regressions take the following form:

$$y_{ij} = \beta_x x_i + \sum_j \beta_j J_{ij} + \sum_f \beta_f F_{fi} + \sum_a \beta_a A_{ai} + \epsilon_{ij}$$

where j indexes individual inventors (or authors). For patents, the fixed effects J_{ij} are a series of individual binary variables $J_{ij} \in \{0,1\}$, where $J_{ij} = 1$ if person j was an inventor of patent i and $J_{ij} = 0$ otherwise. Similarly, for papers, the fixed effects J_{ij} are a series of individual binary variables $J_{ij} \in \{0,1\}$, where $J_{ij} = 1$ if person j was an author of paper i and $J_{ij} = 0$ otherwise. The field and year fixed effects are defined as above.

Regression Methods for Distance and Institutions

Regression analyses to examine the link between institutional type and distance employ fixed-effect ordinary least squares models. These take the following form.

$$D_i = \beta_u u_i + \beta_g g_i + \sum_r \beta_r R_{ri} + \sum_c \beta_c C_{ci} + \sum_n \beta_n N_{ni} + \sum_f \beta_f F_{fi} + \sum_a \beta_a A_{ai} + \epsilon_i$$

where i indexes a specific patent (or paper for the paper regressions). The variables are defined as follows.

Dependent variable: The dependent variable is the integer distance metric as defined in the text.

Predictors of interest: We examine in regression the extent to which university and government laboratory patents (or papers) are nearer or further to the patent-paper boundary, compared to private-sector patents (or papers). Specifically, we define a binary variable $u_i \in \{0,1\}$, where $u_i = 1$ if the institutional setting for the knowledge output is a university, and $u_i = 0$ otherwise. Similarly, we define a binary variable $g_i \in \{0,1\}$, where $g_i = 1$ if the institutional setting for the knowledge output is a government laboratory, and $g_i = 0$ otherwise. The omitted institutional category in the regressions is publicly-traded U.S. firms. Thus the institutional coefficients (β_u and β_g) tell us how the distance for patents (or papers) in these institutional settings differs from those in a private-sector setting.[5]

Fixed effects: To control for other possible influencers of impact and distance in a flexible manner, we include detailed fixed effects as follows.

R_{ri}: These fixed effects account for the number of references the patent (or paper) makes. See detailed definition above.

C_{ci}: These fixed effects account for the number of citations the patent (or paper) receives. The C_{ci} are a series of individual binary variables $C_{ci} \in \{0,1\}$, where $C_{ci} = 1$ if patent (or paper) i receives exactly c citations and $C_{ci} = 0$ otherwise. In practice, we use an individual fixed effect for each integer number of citations received up to 100 and then bin 100 or more into one category.

N_{ni}: These fixed effects account for the number of inventors on the patent or authors on the paper. See detailed definition above.

[5] Note that the regression sample here is restricted to these three institutional types. Patents (or papers) for which the institutional setting could not be determined are not included in the estimation. See "Institutional definitions and matching" under Methods above.

F_{fi}: These fixed effects account for the technological class of the patent or WOS field of the paper. See detailed definition above.

A_{ai}: These fixed effects account for the application year of the patent or publication year of the paper. See detailed definition above.

4. Distance and Citation Counts

When $D > 1$, there can be a natural relationship between distance and reference counts.

- For a patent, the more references the patent makes, the more pathways there are to patent-paper frontier, which can lead to a lower D.
- For a paper, the more citations a paper receives, the more pathways there are to that patent-paper frontier, which can lead to a lower D.

These features of the data do not appear at $D = 1$ because a direct citation from a patent to a paper cannot be constructed from citation behavior between patents or between papers.

Whether it makes sense to account for a relationship between distance and citations counts depends on the question and the analysis. For example, when $D > 1$, a relationship between a paper receiving more citations and smaller D is substantive in the sense that a more impactful paper (among other papers) also opens itself to closer links to future patentable applications. Higher-impact papers, through the multiplicity of downstream work that builds on them, may naturally and meaningfully allow shorter path lengths to a patentable invention. For patents, such a relationship does not follow, because patents connect to papers in our directed graph via backward references to prior patents, not forward citations. For patents, more backwards references may lead to lower D (when $D > 1$) by opening more pathways toward prior work, but this does not imply that the patent will be more cited itself. Nonetheless, and with these nuances in mind, we take several approaches to clarify the results and their robustness to citation counts.

Depending on the question, we confront the linkage between distance and citation counts in three different ways. Our primary method uses fixed effect regressions. Our secondary method focuses on the $D = 1$ case where the linkage does not arise. Our final method uses a null model. This section discusses these methods in turn.

Fixed effects

Our regression models allow us to account in a flexible and non-parametric manner for the number of citations made or received. The citation count fixed effects, by controlling for any average effect from the specific number of citations, mean that these regressions make comparisons across papers or patents that have the same number of references.

The main analysis regarding institutions and distance (Fig. 4) confronts this issue by including fully-flexible controls for citation counts made or received (and shows that the link between institutional type and proximity to the frontier is robust to referencing behavior, as well as numerous other features as discussed above). For example, the regression with fixed effects for each number of citations asks: "Among papers with exactly 21 citations received, are university papers closer to the patent-paper boundary than firm papers?" And similarly for the number of references made, and similarly for the patent regressions. The fixed effects thus neutralize the average effect of any specific number of references on the D of that paper, and thus allow estimation of the institutional differences regarding distance from the patent-paper boundary while accounting for any differences in citations made or received across institutional types.

A second application of citation count fixed effects concerns the ordering of fields in terms of the distance to the patent-paper boundary (Fig. 2). The analyses in the main text categorize fields by their distance from the patent-paper boundary without controlling for potential differences in citation counts across fields. Again, to the extent that being more highly cited substantively affords more pathways toward future patents, fields that are close to the boundary in part due to their importance to downstream work in general is potentially an important part of the story for that field and should not be parsed out. Showing the raw data in the main text also emphasizes transparency. However, adjusting the question slightly, one may still be interested in whether fields are in fact closer or further from the patent-paper boundary due to their differential tendencies to be references in future work. This question can also be analyzed using citation count fixed effects; it turns out that accounting for citations has little effect on the ordering of fields vis-à-vis distance to the boundary.

Fig. S1 presents this finding, showing all science and engineering WOS fields. The x-axis ranks fields over [0,1], ordered by the raw-data D_{mean} for each field. The y-axis presents the ranking after parsing out the effect of citation counts to each paper in each field.

The field ranking that controls for citation counts is determined as follows. In the first step, we run a paper-level regression to predict D_i as a function of a full set of citation count fixed effects. We then take the residual distance measure for each paper from this regression. That is, for all papers with exactly c citations, the regression takes out the average distance among such papers (D^c_{mean}), and each paper is given the residual

$$\widetilde{D_i} = D_i - D^c_{mean}$$

The residual tells us, for each paper, whether it has unusually high or low distance given the number of pathways it has to future work. In the second step, we then average $\widetilde{D_i}$ for each field. This residual mean tells us whether a field is typically closer or further from the patent-paper boundary given the number of citations papers in that field receive.

The y-axis in Fig. S1 presents the rank ordering of fields in their distance from the frontier, using this residual mean. We can see that the rank ordering of fields is very similar whether we account for citation differences across fields (y-axis) or rank fields based on their raw D_{mean}. Thus, for example, mathematics is the farthest field from the frontier either way – its distance from the frontier is not due to fewer citation counts.

$D = 1$

The main text also focuses on citation impact itself (Fig. 3). In this case, for papers, those papers that receive higher citations can be expected to achieve lower distance, per the above discussion, when $D > 1$. Here we of course cannot use fixed effects or other regression methods to control for citations received because citations received is the outcome variable of interest. The main analysis of impact thus focuses on the distinctive impact of $D = 1$ papers. In this case, there is no innate relationship between citation impact and distance, because being linked directly to the other domain is not a construct of referencing within one's own domain.

Note that the impact regressions for patents do not raise this issue, because here the D of a specific patent is determined by the references the patent makes, not the number of citations it receives. In any case, for consistency with the paper presentation, we emphasize the $D = 1$ patent case in the text. For completeness, Fig. S3 shows the home run probability at each integer D. With patents we further examine patent fee renewal payments, rather than citation impact, as an alternative impact measure.

Null Model

Finally, we further explore the relationship between citations and distance using a null model. In this model, we compare the observed D versus an expected D given the number of citations. We build a null model as follows.

Take a focal paper, i, with observed distance D_i. We ask what would happen to D_i if we replaced the citations to the focal paper with randomly selected citing material. This allows us to calculate the expected distance from the frontier for a paper, given a specific number of citations to that paper.

The randomly selected citing entities are drawn from the union of future WOS papers and $D = 1$ patents. Call this set of potential citing entities W_i, and let there be c_i citing entities to the focal paper. In a new random draw of c_i citing entities from the set W_i, we have a set of distance measures among these citing entities (the D of each randomly drawn citing paper, or $D = 0$ for any patent drawn). The distance assigned to paper i will then be the minimum distance in the citing set plus 1. Define this distance assigned to paper i as D_i^r. Then the expected D for the focal paper, $E[D_i]$, will be the expected minimum D_i^r across all possible random draws.

The expected minimum distance can be determined analytically for any number of citing entities, c, using the empirical distribution of D among the papers that might be randomly drawn. Namely, the probability that any particular randomly drawn citing

entity has assigned distance metric D is[6]
$\Pr(D) = \%\ of\ citing\ entities\ in\ W_i\ that\ have\ distance\ metric\ D$

Now, we define $q^{(c)}(D)$ as the probability that with c draws (which represents the number of citations received by an article), the minimum degree is equal to D. To simplify notation, we drop here the subscript i, but keep in mind that these probabilities depend on the publication year of the paper, which determines the set W_i of future citing entities that may cite paper i.

Across c random draws of new citing entities, the probability that the minimum distance for the focal paper will be 1 is

$$q^{(c)}(1) = 1 - \left(1 - Pr(0)\right)^c$$

And the probability for other minimum distances will be

$$q^{(c)}(2) = \binom{c}{1} \Pr(1)\left(1 - Pr(0)\right)^{c-1}$$
$$q^{(c)}(3) = \binom{c}{1} \Pr(2)(1 - Pr(0) - \Pr(1))^{c-1}$$

et cetera. The expected minimum distance for the focal paper is then

$$E[D_i] = \sum_d \mathrm{d} \times q^{(c_i)}(d)$$

where d is the set of positive integers.[7]

Fig. S8 compares the observed distance versus the expected distance for all WOS papers published in 1990. The observed distance is presented as the arithmetic mean of D_i for all papers receiving a given number, c_i, citations. The expected distance is as above.

The figure illustrates two things. First, we see that papers with more citations are closer to the patent-paper boundary, both in the observed data and in the null model. This finding follows naturally as discussed above, and is substantive to the extent that being more highly cited among other papers meaningfully increases the chance that a paper can find a pathway to a patentable application. Second, we see that observed distances are systematically larger than expected distances at all number of citations. This finding illustrates that the random citation network has lower distance between nodes than the actual citation network. Papers in the observed network thus appear to exist in structured knowledge communities that are more weakly connected to other knowledge neighborhoods than random linkages would allow, which acts to extend the lengths of pathways to the patent-paper frontier.

---

[6] Note that this probability is also defined for “disconnected” citing entities for which D is missing and which account for an observable percentage of citing entities in W_i.
[7] If the randomly-drawn citing entity is not connected at any distance to that patent-paper frontier (i.e., it is ‘disconnected’ and has no defined D), then this entity will not affect the expected minimum distance.

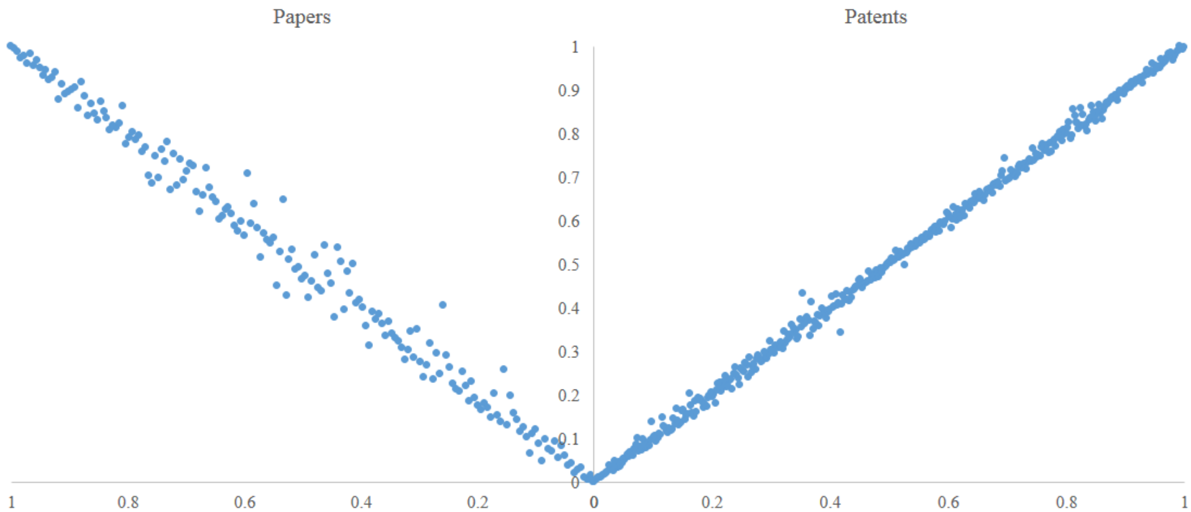


Fig. S1. Field ranks. This figure ranks the mean distance for each field (x-axis) versus the residual mean distance when we account for citation differences across fields (y-axis). See SM text for methods.

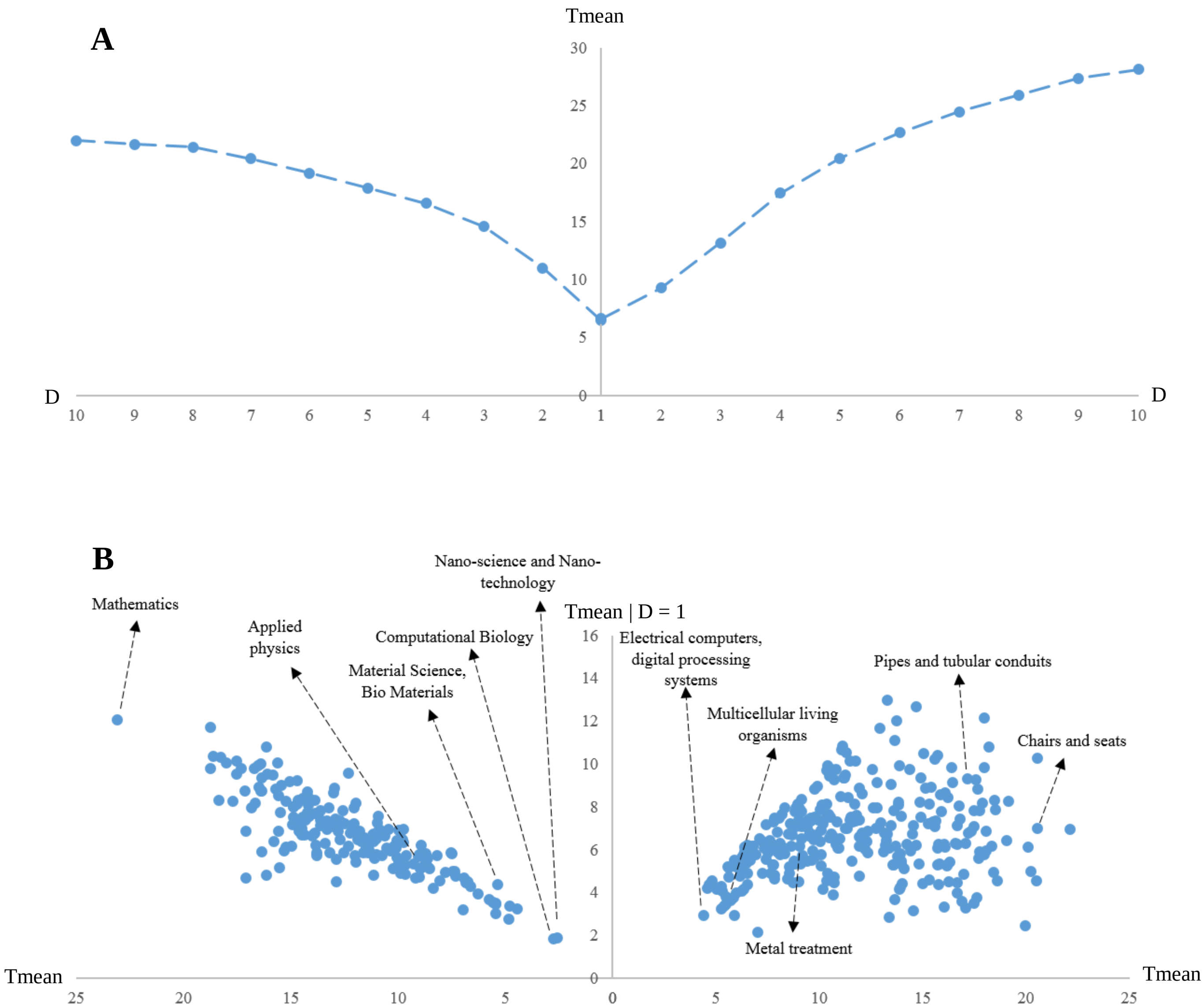


Fig. S2. Time. This figure presents (**A**) time delays by citation distance, averaging across all papers (left) and patents (right). The mean time delay for the field as a whole (x-axis) and average time delay for the field conditional on $D = 1$ (y-axis) are presented for different science fields (left) and different patent classes (right) (**B**).

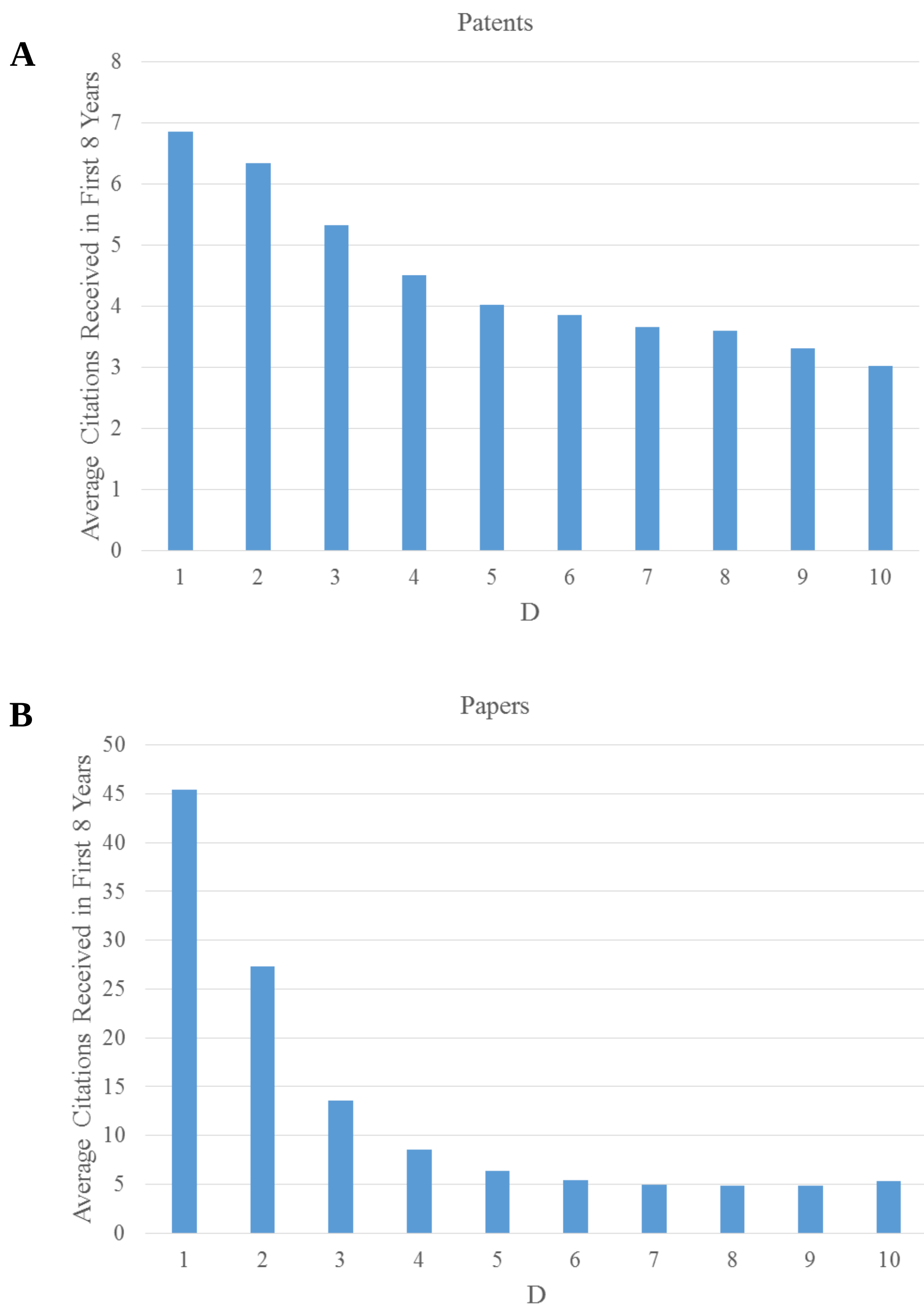


Fig. S3. Impact and Distance at Each Degree, D. This figure presents, for each degree, the mean number of citations received for patents and papers. (**A**) We see that a patent's forward citation impact is greater the more closely its backward citations interact with science. (**B**) For papers, we also see a smooth decay in impact with D, but see further discussion in Section 4 of SM to carefully interpret this finding for papers.

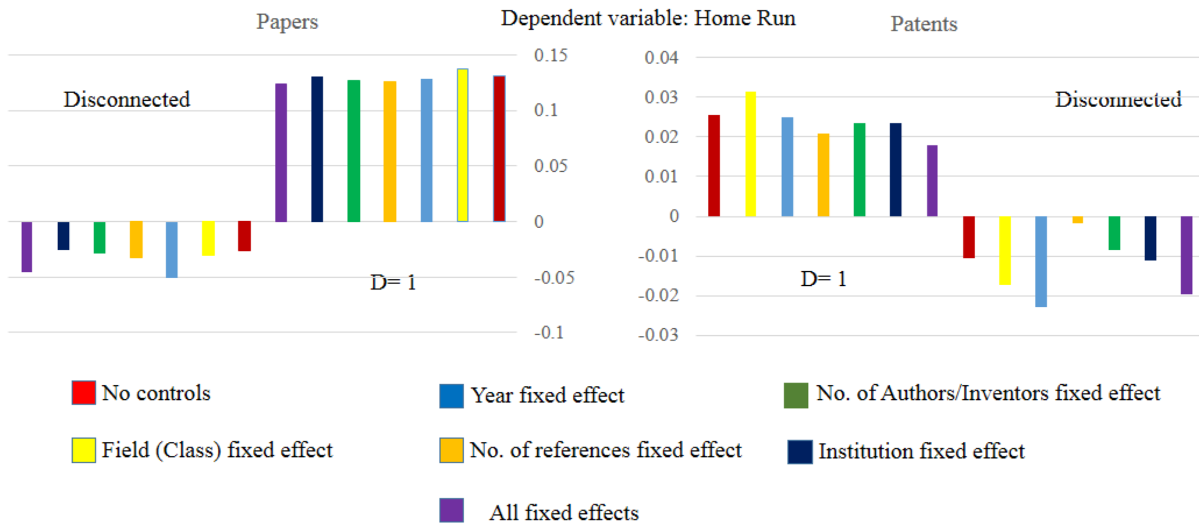


Fig. S4. Regression results show that the impact findings are robust to fixed effects for each year, field/class, institutional setting, number of authors/inventors, and number of references made. The home run definition is being in the upper 5% of citations received in that field and year, for a patent or a research paper. See Tables S3-S4 for the underlying regression results.

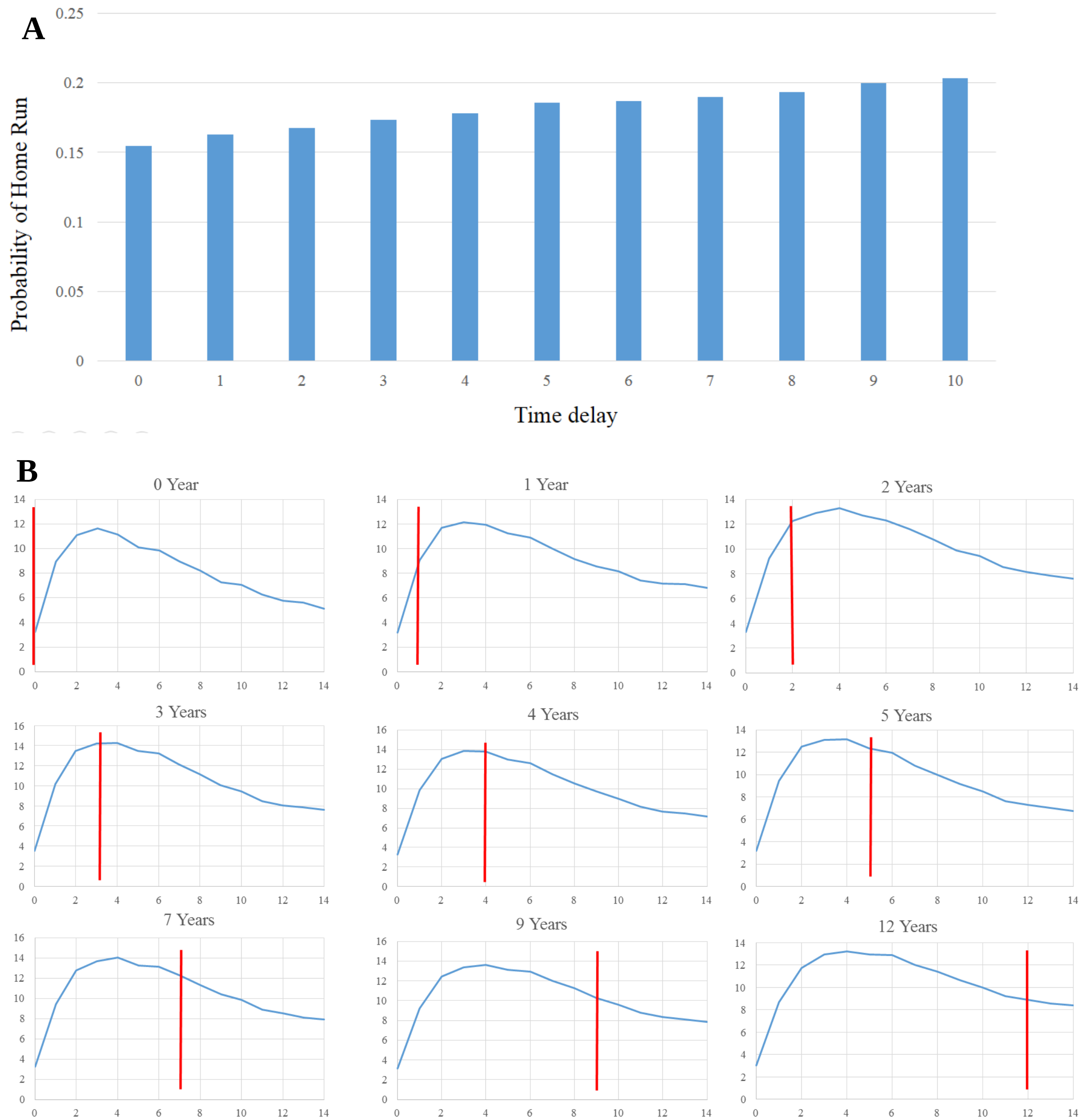


Fig. S5. Timing and paper impact. (**A**) The upper panel shows the home run probability for $D = 1$ papers (y-axis) when grouped by the number of years between patent application and the paper publication (x-axis). The paper's home run probability is three times the background rate of 5% even when a patent cites a paper immediately in the year the paper was published. This finding indicates that $D = 1$ patents do not simply cite already popular, established papers. (**B**) The average citation path over time for $D = 1$ papers, grouping papers by the noted number of years between patent citation and paper publication. The top row of panel B confirms that citations within science largely come after the patent citation for short delay linkages. The lower rows of panel B further indicate that the shape of the citation trajectory within science appears unaffected by the patent citation per se, which appears to rule out a substantial marketing or notoriety effect on scientific papers when cited by a patent.

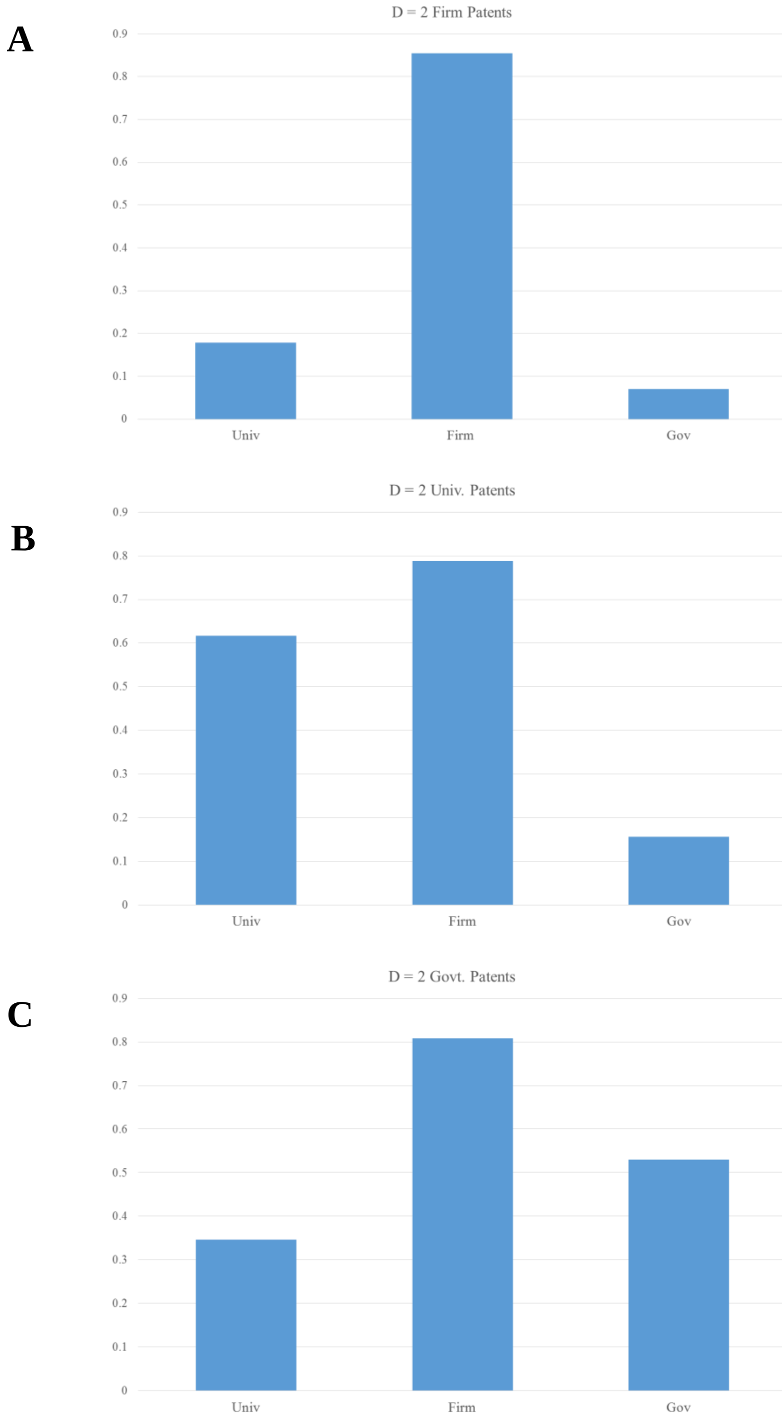


Fig. S6. Institutional pathways to frontier for $D = 2$ patents. (**A**) For $D = 2$ patents from firms, this figure examines the percentage that cite $D = 1$ patents of each institutional type. Percentages (y-axis) add up to more than 100% because a $D = 2$ patent might have multiple pathways to the frontier. Panel (**B**) repeats this analysis for $D = 2$ university patents and (**C**) repeats analysis for $D = 2$ government laboratory patents. We see that all patents tend to go through $D = 1$ firm patents, which follows because most patents are from firms. At the same time, we also see that, along pathways to the frontier, institutions tend to weight upwards their own institutional type in the $D = 1$ patents they cite.

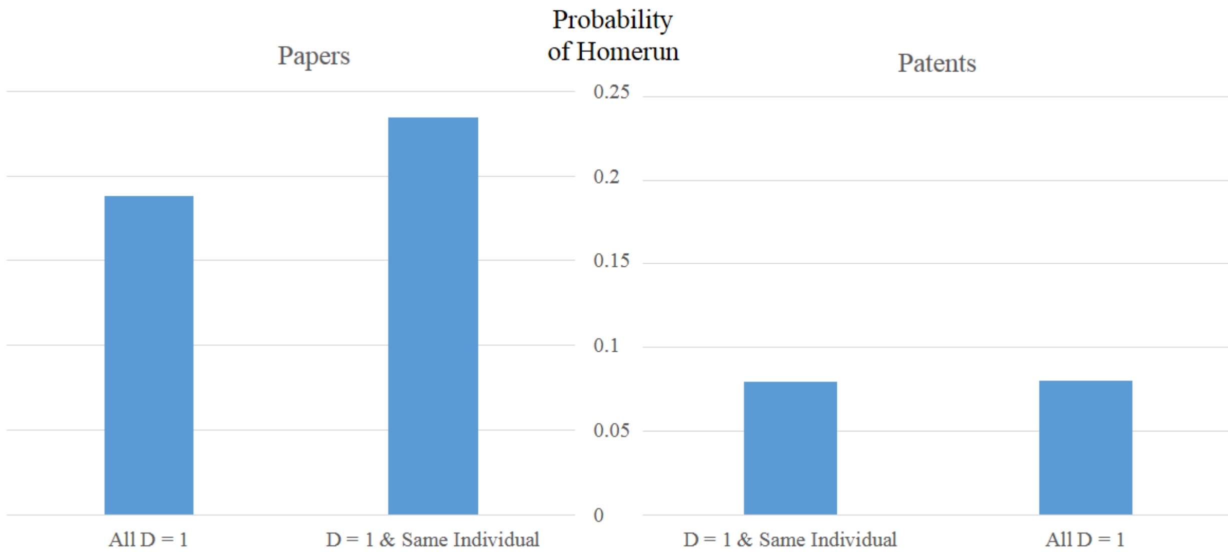


Fig. S7. Impact and Individuals. This figure shows the home run probability for $D = 1$ papers and patents, isolating cases where the same individual is the inventor and the author of the linked patent and paper. These individuals are hitting home-runs at high rates in both domains. The home run is also especially high on the paper side compared to other $D = 1$ papers, while on the patent side the home run rate is very similar to other $D = 1$ patents.

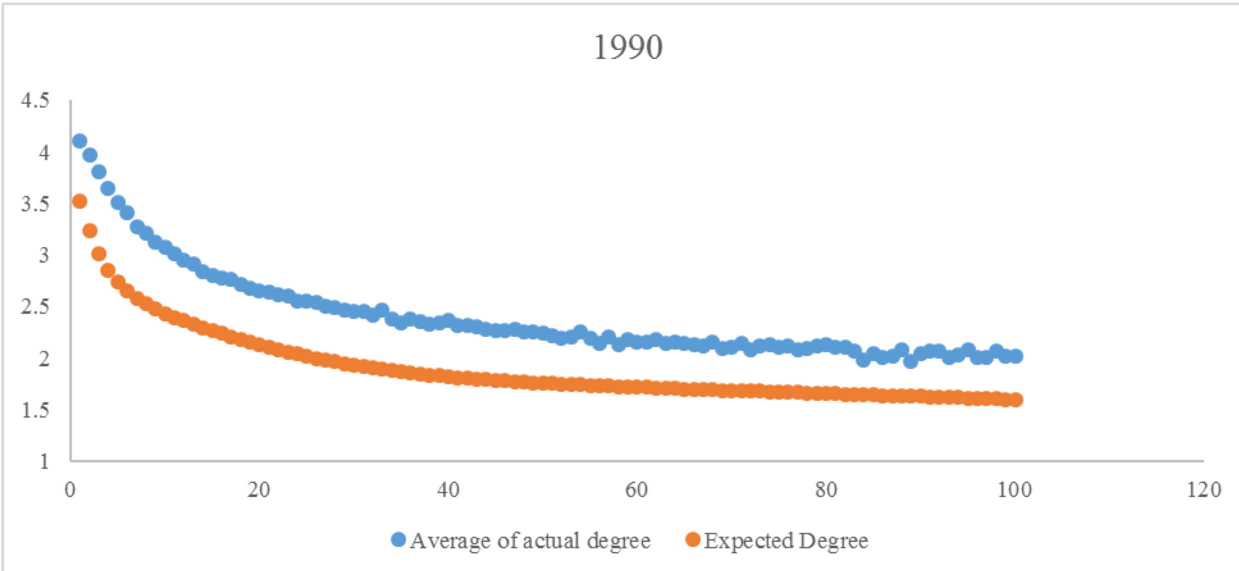
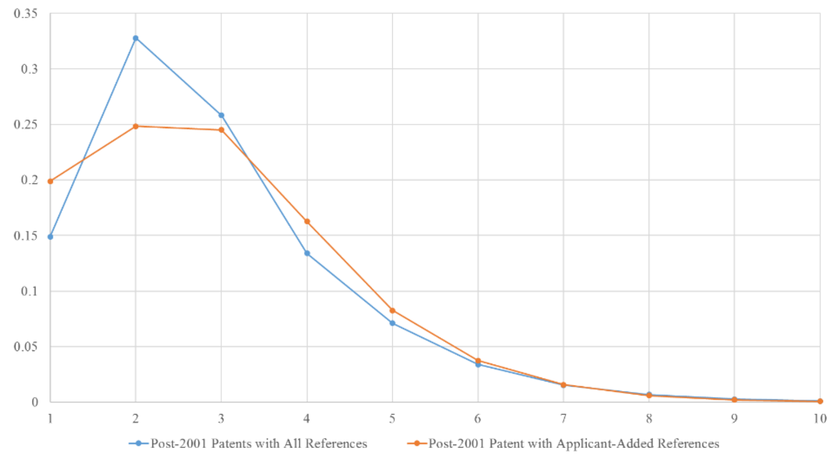


Fig. S8. Null model. This figure presents the observed mean D for each number of citations received (blue) and the expected mean D for each number of citations received assuming randomized citation links (orange). See SM text for methods.

A



B

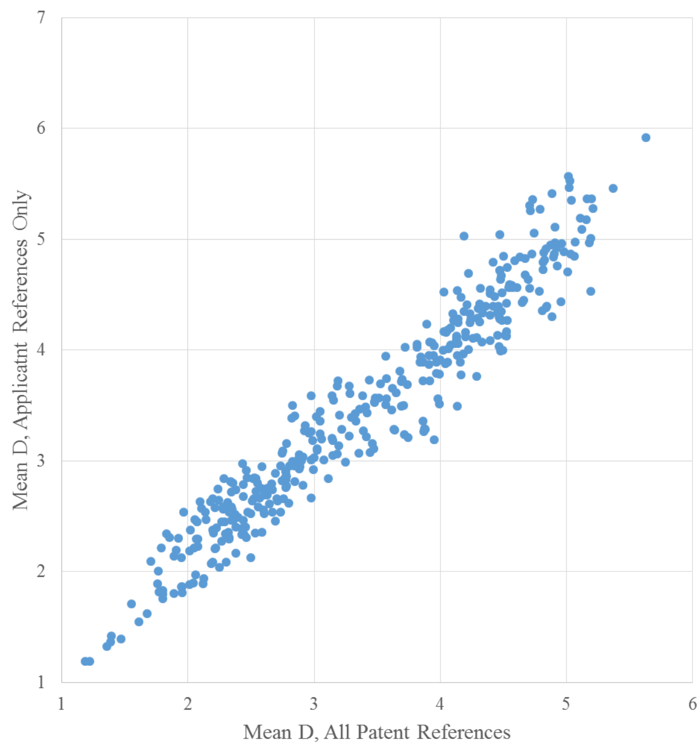


Fig. S9. Applicant-Added Patent References. For post-2001 patents, the database allows one to distinguish patent citations added by patent examiners from those added by applicants. In this figure, we compare the findings for the patent citation network using all references versus a network that only uses applicant-added references. (**A**) Shows some flattening of the peak in the D distribution but the shape is otherwise similar. (**B**) For each patent technology class, the x-axis presents D_{mean} using the network built from all references in each patent while the y-axis presents D_{mean} using the network built from only the applicant-added citations. We see that the technology class ranks are broadly similarly with and without examiner-added citations.

Table S1 – A New Typology for Technology Classes: Mean, mode, and standard deviation of the distance metric and percentage connectivity for all U.S. patent technology classes.

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
800	1.175114	0.221021	1	0.818992	Multicellular Living Organisms and Unmodified Parts Thereof and Related Processes
530	1.297301	0.392269	1	0.794603	Chemistry: Natural Resins or Derivatives; Peptides or Proteins; Lignins or Reaction Products Thereof
435	1.322216	0.428349	1	0.860975	Chemistry: Molecular Biology and Microbiology
505	1.348943	0.383274	1	0.934588	Superconductor Technology: Apparatus, Material, Process
536	1.393703	0.621094	1	0.737323	Organic Compounds -- Part of the Class 532-570 Series
706	1.54019	0.446349	1	0.965108	Data Processing: Artificial Intelligence
514	1.562451	0.812036	1	0.595509	Drug, Bio-Affecting and Body Treating Compositions
552	1.605769	0.746505	1	0.316109	Organic Compounds -- Part of the Class 532-570 Series
546	1.720372	1.031033	1	0.433491	Organic Compounds -- Part of the Class 532-570 Series
372	1.735488	0.586659	1	0.893339	Coherent Light Generators
117	1.741905	0.786892	1	0.832723	Single-Crystal, Oriented-Crystal, and Epitaxy Growth Processes; Non-Coating Apparatus Therefor
424	1.760048	0.907552	1	0.806923	Drug, Bio-Affecting and Body Treating Compositions
436	1.766014	0.759711	1	0.796225	Chemistry: Analytical and Immunological Testing
260	1.780488	1.439619	1	0.488095	Chemistry of Carbon Compounds
549	1.796017	1.11349	1	0.404196	Organic Compounds -- Part of the Class 532-570 Series
548	1.846395	1.363866	1	0.382082	Organic Compounds -- Part of the Class 532-570 Series
707	1.854065	0.385663	2	0.985639	Data Processing: Database and File Management, Data Structures, or Document Processing
540	1.866458	1.39479	1	0.35369	Organic Compounds -- Part of the Class 532-570 Series
544	1.87828	1.328479	1	0.392954	Organic Compounds -- Part of the Class 532-570 Series
554	1.894432	1.198832	1	0.386028	Organic Compounds -- Part of the Class 532-570 Series
704	1.901753	0.511639	2	0.936287	Data Processing: Speech Signal Processing, Linguistics, Language Translation, and Audio Compression/Decompression
712	1.948629	0.577034	2	0.940221	Electrical Computers and Digital Processing Systems: Processing Architectures and Instruction Processing (e.g., Processors)
564	1.951075	1.363404	1	0.374102	Organic Compounds -- Part of the Class 532-570 Series
709	1.968096	0.391028	2	0.988196	Electrical Computers and Digital Processing Systems: Multiple Computer or Process Coordinating
570	2.002865	1.143258	1	0.352882	Organic Compounds -- Part of the Class 532-570 Series
607	2.014299	0.73059	2	0.874205	Surgery: Light, Thermal, and Electrical Application
382	2.020667	0.67071	2	0.957274	Image Analysis
560	2.02649	1.482742	1	0.332188	Organic Compounds -- Part of the Class 532-570 Series
562	2.03675	1.49188	1	0.33014	Organic Compounds -- Part of the Class 532-570 Series
600	2.04591	0.78718	2	0.861807	Surgery
385	2.06043	0.756556	2	0.943234	Optical Waveguides
708	2.070423	0.738771	2	0.816664	Electrical Computers: Arithmetic Processing and Calculating
438	2.071271	0.591287	2	0.958071	Semiconductor Device Manufacturing: Process
623	2.07854	0.670164	2	0.892124	Prosthesis (i.e., Artificial Body Members), Parts Thereof, or Aids and Accessories Therefor
558	2.079504	1.677851	1	0.294522	Organic Compounds -- Part of the Class 532-570 Series
370	2.106992	0.559557	2	0.938243	Multiplex Communications

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
714	2.137949	0.599837	2	0.9326	Error Detection/Correction and Fault Detection/Recovery
375	2.139796	0.731083	2	0.915033	Pulse or Digital Communications
556	2.140344	1.437301	1	0.543908	Organic Compounds -- Part of the Class 532-570 Series
127	2.157593	1.296081	2	0.43462	Sugar, Starch, and Carbohydrates
356	2.171978	1.0256	2	0.763407	Optics: Measuring and Testing
568	2.178053	1.678927	1	0.361082	Organic Compounds -- Part of the Class 532-570 Series
423	2.195493	1.15998	2	0.480954	Chemistry of Inorganic Compounds
518	2.195506	1.208969	2	0.489549	Chemistry: Fischer-Tropsch Processes; or Purification or Recovery of Products Thereof
711	2.195665	0.551217	2	0.947604	Electrical Computers and Digital Processing Systems: Memory
257	2.206684	0.827155	2	0.92005	Active Solid-State Devices (e.g., Transistors, Solid-State Diodes)
512	2.210169	1.650744	1	0.327778	Perfume Compositions
216	2.220155	1.17946	2	0.807442	Etching a Substrate: Processes
702	2.230358	0.923237	2	0.916401	Data Processing: Measuring, Calibrating, or Testing
705	2.237093	0.682755	2	0.933495	Data Processing: Financial, Business Practice, Management, or Cost/Price Determination
136	2.241316	1.164801	2	0.777817	Batteries: Thermoelectric and Photoelectric
380	2.260653	0.872029	2	0.883242	Cryptography
204	2.27239	1.380588	2	0.644018	Chemistry: Electrical and Wave Energy
341	2.274997	0.95966	2	0.8229	Coded Data Generation or Conversion
326	2.280042	0.716282	2	0.924535	Electronic Digital Logic Circuitry
713	2.28436	0.567348	2	0.972782	Electrical Computers and Digital Processing Systems: Support
606	2.294965	0.912996	2	0.85179	Surgery
252	2.295794	1.29058	2	0.64531	Compositions
330	2.297147	0.955432	2	0.684265	Amplifiers
367	2.306974	1.232666	2	0.574187	Communications, Electrical: Acoustic Wave Systems and Devices
585	2.3103	1.257605	2	0.464387	Chemistry of Hydrocarbon Compounds
378	2.312279	1.141794	2	0.702416	X-Ray or Gamma Ray Systems or Devices
504	2.312295	1.463947	2	0.337997	Plant Protecting and Regulating Compositions
333	2.323264	1.068898	2	0.694242	Wave Transmission Lines and Networks
250	2.333152	1.238335	2	0.741571	Radiant Energy
342	2.341514	0.915302	2	0.675589	Communications: Directive Radio Wave Systems and Devices (e.g., Radar, Radio Navigation)
710	2.342099	0.496668	2	0.942519	Electrical Computers and Digital Data Processing Systems: Input/Output
349	2.343737	0.748129	2	0.938112	Liquid Crystal Cells, Elements and Systems
345	2.355601	0.736356	2	0.929229	Computer Graphics Processing, Operator Interface Processing, and Selective Visual Display Systems
374	2.35859	1.194347	2	0.67047	Thermal Measuring and Testing
588	2.360444	0.89411	2	0.712311	Hazardous or Toxic Waste Destruction or Containment
502	2.367155	1.414365	2	0.508629	Catalyst, Solid Sorbent, or Support Therefor: Product or Process of Making
365	2.377032	0.853496	2	0.902705	Static Information Storage and Retrieval
501	2.388352	1.047848	2	0.711548	Compositions: Ceramic
426	2.396419	1.572398	2	0.548394	Food or Edible Material: Processes, Compositions, and Products
205	2.399323	1.623254	2	0.484653	Electrolysis: Processes, Compositions Used Therein, and Methods of Preparing the Compositions

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
420	2.407789	1.327408	2	0.311786	Alloys or Metallic Compositions
148	2.420596	1.65962	2	0.464031	Metal Treatment
324	2.424928	1.100078	2	0.774617	Electricity: Measuring and Testing
455	2.434308	0.656267	2	0.895207	Telecommunications
526	2.435057	1.835474	2	0.533704	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
527	2.436975	1.960314	1	0.477912	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
379	2.437166	0.833557	2	0.799147	Telephonic Communications
507	2.438017	1.070538	2	0.528384	Earth Boring, Well Treating, and Oil Field Chemistry
700	2.462103	0.938924	2	0.877077	Data Processing: Generic Control Systems or Specific Applications
327	2.464759	0.873309	2	0.817436	Miscellaneous Active Electrical Nonlinear Devices, Circuits, and Systems
95	2.471297	1.329417	2	0.650371	Gas Separation: Processes
NA	2.471429	1.761565	2	0.804598	
427	2.484185	1.607108	2	0.602675	Coating Processes
359	2.498096	1.758606	2	0.752874	Optics: Systems (Including Communication) and Elements
343	2.512425	1.084135	2	0.700268	Communications: Radio Wave Antennas
71	2.512871	1.90528	2	0.381997	Chemistry: Fertilizers
522	2.515225	1.42151	2	0.677113	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
386	2.526687	0.705254	2	0.853147	Television Signal Processing for Dynamic Recording or Reproducing
331	2.531557	1.420241	2	0.681113	Oscillators
51	2.534729	1.177433	2	0.685137	Abrasive Tool Making Process, Material, or Composition
348	2.535461	0.885106	2	0.819967	Television
534	2.547408	2.283151	1	0.297592	Organic Compounds -- Part of the Class 532-570 Series
128	2.553712	1.509109	2	0.708701	Surgery
445	2.558483	1.87897	2	0.646179	Electric Lamp or Space Discharge Component or Device Manufacturing
332	2.565385	1.295725	2	0.625	Modulators
323	2.585055	0.971473	2	0.769739	Electricity: Power Supply or Regulation Systems
65	2.590652	1.508407	2	0.502285	Glass Manufacturing
516	2.591973	1.659601	2	0.550645	Colloid Systems and Wetting Agents; Subcombinations Thereof; Processes Of
422	2.597975	1.496437	2	0.636879	Chemical Apparatus and Process Disinfecting, Deodorizing, Preserving, or Sterilizing
334	2.607143	0.73852	2	0.191781	Tuners
196	2.608696	1.10775	2	0.176923	Mineral Oils: Apparatus
419	2.615445	1.299075	2	0.634339	Powder Metallurgy Processes
291	2.625	1.734375	3	0.242424	Track Sanders
510	2.625794	1.061209	2	0.695186	Cleaning Compositions for Solid Surfaces, Auxiliary Compositions Therefor, or Processes of Preparing the Compositions
75	2.632	1.701667	2	0.406564	Specialized Metallurgical Processes, Compositions for Use Therein, Consolidated Metal Powder Compositions, and Loose Metal Particulate Mixtures
162	2.656138	1.723561	2	0.547436	Paper Making and Fiber Liberation
381	2.662545	1.210124	2	0.723589	Electrical Audio Signal Processing Systems and Devices
604	2.665856	1.271812	2	0.754056	Surgery
23	2.666667	2.188889	2	0.337079	Chemistry: Physical Processes
429	2.682677	1.690645	2	0.660034	Chemistry: Electrical Current Producing Apparatus, Product, and Process

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
528	2.685146	2.212824	2	0.468968	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
210	2.697241	1.650995	2	0.556124	Liquid Purification or Separation
159	2.709924	1.427306	2	0.253385	Concentrating Evaporators
149	2.717252	1.199607	2	0.380085	Explosive and Thermic Compositions or Charges
1	2.722222	1.533951	2	0.9	** Classification Undetermined **
363	2.743364	1.033804	3	0.717674	Electric Power Conversion Systems
208	2.743845	1.387797	2	0.428495	Mineral Oils: Processes and Products
203	2.744741	1.51689	2	0.362297	Distillation: Processes, Separatory
376	2.745477	1.724448	2	0.411459	Induced Nuclear Reactions: Processes, Systems, and Elements
360	2.754017	1.404029	2	0.705424	Dynamic Magnetic Information Storage or Retrieval
329	2.755556	1.703883	2	0.531686	Demodulators
73	2.761501	1.614618	2	0.599299	Measuring and Testing
525	2.767359	2.010653	2	0.488415	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
48	2.770925	1.590697	2	0.325448	Gas: Heating and Illuminating
351	2.771874	1.926704	2	0.527368	Optics: Eye Examining, Vision Testing and Correcting
118	2.774626	1.731499	2	0.592752	Coating Apparatus
369	2.781043	1.001925	3	0.865226	Dynamic Information Storage or Retrieval
494	2.800745	1.455642	2	0.39083	Imperforate Bowl: Centrifugal Separators
433	2.801966	1.50142	2	0.541688	Dentistry
235	2.823529	0.928585	3	0.789479	Registers
175	2.832048	1.389394	2	0.483472	Boring or Penetrating the Earth
463	2.832151	0.829983	3	0.826172	Amusement Devices: Games
338	2.832182	1.548495	2	0.515302	Electrical Resistors
166	2.843605	1.735715	2	0.530373	Wells
353	2.843943	0.8302	3	0.728502	Optics: Image Projectors
313	2.848886	1.923822	2	0.602686	Electric Lamp and Discharge Devices
508	2.860978	1.672907	2	0.417869	Solid Anti-Friction Devices, Materials Therefor, Lubricant or Separate Compositions for Moving Solid Surfaces, and Miscellaneous Mineral Oil Compositions
106	2.867481	1.70914	2	0.511152	Compositions: Coating or Plastic
523	2.876069	1.828561	2	0.497999	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
377	2.878268	1.147763	3	0.534731	Electrical Pulse Counters, Pulse Dividers, or Shift Registers: Circuits and Systems
428	2.879076	1.941092	2	0.626129	Stock Material or Miscellaneous Articles
352	2.881633	1.467622	3	0.255208	Optics: Motion Pictures
8	2.8921	2.010708	2	0.400463	Bleaching and Dyeing; Fluid Treatment and Chemical Modification of Textiles and Fibers
434	2.90827	1.889431	2	0.523178	Education and Demonstration
315	2.918019	1.420227	3	0.605802	Electric Lamp and Discharge Devices: Systems
201	2.931034	1.20214	3	0.121339	Distillation: Processes, Thermolytic
358	2.931576	1.013178	3	0.898132	Facsimile and Static Presentation Processing
131	2.949398	2.072138	2	0.243759	Tobacco
318	2.961104	1.3377	3	0.651396	Electricity: Motive Power Systems
178	2.967177	1.611614	3	0.499454	Telegraphy

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
228	2.975948	1.348964	3	0.581661	Metal Fusion Bonding
701	2.977181	1.2299	3	0.917627	Data Processing: Vehicles, Navigation, and Relative Location
430	2.980777	1.784392	3	0.694869	Radiation Imagery Chemistry: Process, Composition, or Product Thereof
521	2.987399	1.930374	3	0.448616	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
44	2.998458	1.807245	3	0.410703	Fuel and Related Compositions
524	3.015944	1.998646	3	0.474148	Synthetic Resins or Natural Rubbers -- Part of the Class 520 Series
96	3.021657	1.662769	3	0.494247	Gas Separation: Apparatus
340	3.022138	1.233827	3	0.718952	Communications: Electrical
134	3.045502	1.979315	2	0.588844	Cleaning and Liquid Contact with Solids
355	3.046447	1.702665	3	0.595616	Photocopying
322	3.048227	1.543773	3	0.571313	Electricity: Single Generator Systems
388	3.060127	1.214739	3	0.392547	Electricity: Motor Control Systems
264	3.070119	2.056779	2	0.543079	Plastic and Nonmetallic Article Shaping or Treating: Processes
601	3.102228	2.619956	2	0.556731	Surgery: Kinesitherapy
219	3.135265	1.687445	3	0.592233	Electric Heating
290	3.135843	1.832611	3	0.550947	Prime-Mover Dynamo Plants
442	3.140203	1.993012	3	0.547388	Web or Sheet Containing Structurally Defined Element or Component (428/221)
86	3.151786	2.003747	3	0.23382	Ammunition and Explosive-Charge Making
451	3.15623	1.986992	3	0.527662	Abrading
244	3.178449	2.223915	2	0.416152	Aeronautics
347	3.184677	1.346314	3	0.817894	Incremental Printing of Symbolic Information
361	3.191165	1.454121	3	0.701816	Electricity: Electrical Systems and Devices
346	3.191589	2.042733	3	0.308802	Recorders
310	3.19559	2.138665	3	0.637064	Electrical Generator or Motor Structure
602	3.214477	1.740416	3	0.57473	Surgery: Splint, Brace, or Bandage
89	3.247972	2.047432	3	0.236763	Ordnance
202	3.261708	1.8406	3	0.282271	Distillation: Apparatus
320	3.27776	1.380081	3	0.750642	Electricity: Battery or Capacitor Charging or Discharging
307	3.282966	1.524086	3	0.669529	Electrical Transmission or Interconnection Systems
336	3.295356	2.192155	3	0.497832	Inductor Devices
156	3.325683	1.947308	3	0.490886	Adhesive Bonding and Miscellaneous Chemical Manufacture
234	3.333333	0.622222	3	0.294118	Selective Cutting (e.g., Punching)
76	3.335404	1.899927	3	0.199504	Metal Tools and Implements, Making
164	3.349802	2.765978	2	0.296774	Metal Founding
29	3.360501	2.266597	3	0.455759	Metal Working
209	3.386613	2.375006	3	0.341813	Classifying, Separating, and Assorting Solids
246	3.389058	1.410944	3	0.408696	Railway Switches and Signals
47	3.40702	2.779532	4	0.390478	Plant Husbandry
102	3.407632	1.59701	3	0.310105	Ammunition and Explosives
181	3.407903	2.166138	3	0.429392	Acoustics
392	3.41771	2.209239	3	0.492728	Electric Resistance Heating Devices
295	3.421053	2.875346	2	0.223529	Railway Wheels and Axles

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
174	3.435782	1.840733	3	0.632111	Electricity: Conductors and Insulators
266	3.45302	2.522961	3	0.174985	Metallurgical Apparatus
373	3.473529	3.178711	3	0.226516	Industrial Electric Heating Furnaces
110	3.47528	2.323219	3	0.475102	Furnaces
34	3.481234	2.630799	3	0.405068	Drying and Gas or Vapor Contact with Solids
492	3.488722	1.447868	3	0.472189	Roll or Roller
335	3.49595	3.197101	3	0.393661	Electricity: Magnetically Operated Switches, Magnets, and Electromagnets
177	3.525389	1.905314	3	0.372731	Weighing Scales
60	3.56543	2.423047	3	0.483332	Power Plants
165	3.56544	1.955417	3	0.437646	Heat Exchange
432	3.569154	2.32283	3	0.305936	Heating
405	3.570769	2.681915	3	0.341817	Hydraulic and Earth Engineering
368	3.609868	1.885297	3	0.328791	Horology: Time Measuring Systems or Devices
366	3.611167	2.34227	3	0.367128	Agitating
261	3.633663	2.638075	3	0.343246	Gas and Liquid Contact Apparatus
111	3.635036	2.728116	3	0.41673	Planting
87	3.638298	2.784065	3	0.370079	Textiles: Braiding, Netting, and Lace Making
62	3.649486	2.589288	3	0.508508	Refrigeration
283	3.679245	1.678248	3	0.558483	Printed Matter
299	3.68254	3.664298	3	0.135019	Mining or In Situ Disintegration of Hard Material
416	3.693938	2.239002	3	0.417662	Fluid Reaction Surfaces (i.e., Impellers)
225	3.701031	3.130549	3	0.287549	Severing by Tearing or Breaking
186	3.701613	0.999675	3	0.504065	Merchandising
169	3.70229	2.875745	3	0.274059	Fire Extinguishers
194	3.725962	1.692531	3	0.386378	Check-Actuated Control Mechanisms
84	3.737492	2.734284	3	0.40365	Music
227	3.744898	3.464608	2	0.375087	Elongated-Member-Driving Apparatus
415	3.816134	2.422589	4	0.427462	Rotary Kinetic Fluid Motors or Pumps
55	3.817762	1.790319	4	0.49557	Gas Separation
417	3.838311	2.458512	3	0.41172	Pumps
171	3.842105	3.711911	3	0.075099	Unearthing Plants or Buried Objects
33	3.845835	2.457242	3	0.342929	Geometrical Instruments
503	3.852399	2.676861	3	0.296283	Record Receiver Having Plural Interactive Leaves or a Colorless Color Former, Method of Use, or Developer Therefor
14	3.865217	2.525312	3	0.244941	Bridges
122	3.866292	2.275381	3	0.364008	Liquid Heaters and Vaporizers
452	3.868644	3.177661	3	0.179878	Butchering
168	3.869565	1.374291	3	0.2	Farriery
413	3.869565	2.374291	3	0.142415	Sheet Metal Container Making
140	3.869565	3.591682	4	0.094553	Wireworking
407	3.874539	2.570975	3	0.341956	Cutters, for Shaping
400	3.897408	1.842534	4	0.505674	Typewriting Machines

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
119	3.898658	2.567582	4	0.375268	Animal Husbandry
289	3.90625	2.084961	4	0.166667	Knots and Knot Tying
218	3.907186	3.790787	3	0.190966	High-Voltage Switches with Arc Preventing or Extinguishing Devices
104	3.910331	2.413012	3	0.207861	Railways
425	3.916791	2.604909	4	0.38135	Plastic Article or Earthenware Shaping or Treating: Apparatus
138	3.933559	1.837928	4	0.438952	Pipes and Tubular Conduits
236	3.942641	2.131991	4	0.393862	Automatic Temperature and Humidity Regulation
460	3.953125	3.034261	4	0.225352	Crop Threshing or Separating
404	3.955642	1.723324	4	0.381079	Road Structure, Process, or Apparatus
431	3.965164	2.898377	4	0.377855	Combustion
19	3.972851	1.795643	4	0.25286	Textiles: Fiber Preparation
249	3.984756	2.685743	4	0.191142	Static Molds
43	3.988145	3.123748	3	0.241518	Fishing, Trapping, and Vermin Destroying
241	3.991228	3.167939	3	0.237805	Solid Material Comminution or Disintegration
28	4.002237	3.648765	3	0.315678	Textiles: Manufacturing
384	4.024414	2.411563	4	0.294267	Bearings
273	4.028936	1.858737	4	0.296268	Amusement Devices: Games
68	4.040964	3.085069	4	0.215696	Textiles: Fluid Treating Apparatus
187	4.051038	2.648779	4	0.394809	Elevator, Industrial Lift Truck, or Stationary Lift for Vehicle
239	4.057437	2.700087	4	0.379041	Fluid Sprinkling, Spraying, and Diffusing
12	4.060241	4.080708	2	0.163708	Boot and Shoe Making
125	4.065934	2.501147	3	0.353398	Stone Working
277	4.078769	2.402125	4	0.361538	Seal for a Joint or Juncture
477	4.086399	1.434388	4	0.757968	Interrelated Power Delivery Controls, Including Engine Control
362	4.104522	2.331652	3	0.584903	Illumination
482	4.107482	1.750752	4	0.553168	Exercise Devices
191	4.117117	3.238536	4	0.226069	Electricity: Transmission To Vehicles
116	4.119545	2.276032	4	0.267921	Signals and Indicators
126	4.121341	3.214266	3	0.277661	Stoves and Furnaces
414	4.132334	2.802338	3	0.286997	Material or Article Handling
82	4.132653	2.841587	4	0.238675	Turning
399	4.136417	1.613027	4	0.70389	Electrophotography
251	4.136624	2.46701	4	0.313484	Valves and Valve Actuation
409	4.143406	2.68366	3	0.300616	Gear Cutting, Milling, or Planing
5	4.144492	2.9023	3	0.387804	Beds
212	4.159091	2.870145	4	0.195382	Traversing Hoists
101	4.166075	2.495035	3	0.457125	Printing
472	4.175793	3.23999	3	0.305996	Amusement Devices
72	4.177816	3.030099	4	0.201671	Metal Deforming
293	4.180451	2.177964	4	0.39662	Vehicle Fenders
473	4.189194	2.002561	4	0.417732	Games Using Tangible Projectile
406	4.19375	2.518711	4	0.189237	Conveyors: Fluid Current

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
446	4.200978	2.384547	4	0.326886	Amusement Devices: Toys
141	4.215061	2.633959	4	0.460558	Fluent Material Handling, with Receiver or Receiver Coacting Means
439	4.222651	2.029756	4	0.608336	Electrical Connectors
337	4.225473	3.806305	3	0.258222	Electricity: Electrothermally or Thermally Actuated Switches
139	4.229465	2.510579	3	0.221484	Textiles: Weaving
137	4.233786	2.787005	4	0.296376	Fluid Handling
152	4.243519	2.15908	4	0.303384	Resilient Tires and Wheels
462	4.245283	2.562478	3	0.24537	Books, Strips, and Leaves for Manifolding
470	4.245614	3.799323	3	0.156164	Threaded, Headed Fastener, or Washer Making: Process and Apparatus
57	4.259972	2.530764	4	0.170218	Textiles: Spinning, Twisting, and Twining
453	4.269841	2.435122	4	0.208955	Coin Handling
408	4.287902	3.071783	3	0.205669	Cutting by Use of Rotating Axially Moving Tool
100	4.291492	3.690486	4	0.226397	Presses
83	4.300875	2.442297	4	0.22742	Cutting
294	4.314534	2.79044	4	0.185327	Handling: Hand and Hoist-Line Implements
396	4.31553	2.522401	4	0.453856	Photography
180	4.317217	2.732882	4	0.420527	Motor Vehicles
188	4.32801	3.105759	4	0.312716	Brakes
222	4.336061	2.198392	4	0.388952	Dispensing
269	4.337646	2.588453	4	0.267289	Work Holders
26	4.361446	1.91755	4	0.166333	Textiles: Cloth Finishing
2	4.395052	2.859896	4	0.407568	Apparel
450	4.395918	1.831004	4	0.382813	Foundation Garments
105	4.397119	2.40814	4	0.223243	Railway Rolling Stock
109	4.398374	1.703087	4	0.242604	Safes, Bank Protection, or a Related Device
493	4.420569	2.210633	4	0.298356	Manufacturing Container or Tube From Paper; or Other Manufacturing From a Sheet or Web
66	4.422062	2.689969	4	0.157418	Textiles: Knitting
221	4.424157	2.468967	4	0.314766	Article Dispensing
226	4.456233	2.794504	4	0.224271	Advancing Material of Indeterminate Length
132	4.457143	3.089433	4	0.345205	Toilet
114	4.459477	3.925448	3	0.199136	Ships
15	4.460214	2.424097	4	0.257427	Brushing, Scrubbing, and General Cleaning
285	4.466914	2.125485	4	0.346649	Pipe Joints or Couplings
215	4.467074	1.7832	4	0.349135	Bottles and Jars
59	4.467742	1.700572	5	0.093514	Chain, Staple, and Horseshoe Making
237	4.470588	2.366782	5	0.307034	Heating Systems
383	4.471944	2.243201	4	0.447734	Flexible Bags
123	4.472974	2.1713	4	0.552693	Internal-Combustion Engines
42	4.477352	2.586305	4	0.244394	Firearms
184	4.484375	2.531006	4	0.313725	Lubrication
27	4.484848	1.934619	4	0.419847	Undertaking

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
52	4.485391	2.708108	4	0.321736	Static Structures (e.g., Buildings)
173	4.490625	2.890537	4	0.270042	Tool Driving or Impacting
267	4.49113	2.931527	4	0.297005	Spring Devices
254	4.494828	3.198249	4	0.162784	Implements or Apparatus for Applying Pushing or Pulling Force
172	4.51046	3.053238	3	0.20515	Earth Working
53	4.526302	2.850718	4	0.355744	Package Making
99	4.526316	2.5384	4	0.387354	Foods and Beverages: Apparatus
441	4.527964	2.866668	5	0.211848	Buoys, Rafts, and Aquatic Devices
305	4.533333	2.064274	4	0.237805	Wheel Substitutes for Land Vehicles
483	4.534296	2.718138	4	0.33134	Tool Changing
454	4.543614	2.726291	4	0.352844	Ventilation
37	4.545337	2.763489	4	0.279913	Excavating
157	4.566667	2.445556	5	0.073892	Wheelwright Machines
401	4.568273	2.873853	4	0.261417	Coating Implements with Material Supply
40	4.569264	2.430267	4	0.309548	Card, Picture, or Sign Exhibiting
303	4.59027	2.449377	4	0.500095	Fluid-Pressure and Analogous Brake Systems
144	4.610338	2.245778	4	0.155487	Woodworking
63	4.624309	2.52184	5	0.190928	Jewelry
412	4.630303	2.729991	4	0.317919	Bookbinding: Process and Apparatus
56	4.643212	2.828828	4	0.213465	Harvesters
92	4.664444	2.367402	4	0.291262	Expansible Chamber Devices
206	4.673001	2.706532	4	0.379576	Special Receptacle or Package
7	4.674699	1.701408	4	0.211735	Compound Tools
476	4.681992	3.971668	3	0.504836	Friction Gear Transmission Systems or Components
280	4.708079	2.35531	4	0.396586	Land Vehicles
124	4.710497	2.210111	4	0.379137	Mechanical Guns and Projectors
200	4.717956	3.394904	4	0.359052	Electricity: Circuit Makers and Breakers
301	4.730878	2.089047	5	0.361866	Land Vehicles: Wheels and Axles
160	4.731463	1.971976	4	0.376699	Flexible or Portable Closure, Partition, or Panel
16	4.747026	2.485568	5	0.265418	Miscellaneous Hardware
185	4.772727	1.448347	5	0.177419	Motors: Spring, Weight, or Animal Powered
30	4.785998	3.108539	4	0.17231	Cutlery
193	4.793478	2.816044	5	0.155932	Conveyors, Chutes, Skids, Guides, and Ways
296	4.795501	2.133913	5	0.475618	Land Vehicles: Bodies and Tops
464	4.811494	3.649523	4	0.282927	Rotary Shafts, Gudgeons, Housings, and Flexible Couplings for Rotary Shafts
403	4.811881	2.978589	5	0.271505	Joints and Connections
242	4.816553	2.675699	4	0.234382	Winding, Tensioning, or Guiding
135	4.820513	2.382089	5	0.311606	Tent, Canopy, Umbrella, or Cane
248	4.826147	2.733422	4	0.335323	Supports
418	4.836791	3.83505	5	0.167944	Rotary Expansible Chamber Devices
220	4.837478	2.617311	4	0.328364	Receptacles
256	4.846868	2.988151	4	0.315058	Fences

Table S1 - Continued

nClass	Dmean	DVariance	Dmode	Connectivity	Class Name
198	4.875163	2.567422	5	0.28294	Conveyors: Power-Driven
270	4.885246	2.012167	5	0.438849	Sheet-Material Associating
411	4.887841	3.156183	4	0.198957	Expanded, Threaded, Driven, Headed, Tool-Deformed, or Locked-Threaded Fastener
474	4.9017	2.882429	4	0.319707	Endless Belt Power Transmission Systems or Components
24	4.905827	2.797488	4	0.24684	Buckles, Buttons, Clasps, Etc.
112	4.909315	2.03681	5	0.311551	Sewing
440	4.910472	3.071121	4	0.318645	Marine Propulsion
49	4.921875	2.447021	5	0.276942	Movable or Removable Closures
74	4.925116	2.93886	5	0.344843	Machine Element or Mechanism
312	4.937207	2.399226	4	0.323647	Supports: Cabinet Structure
38	4.960474	2.662469	4	0.305187	Textiles: Ironing or Smoothing
211	4.966134	2.714751	4	0.341252	Supports: Racks
232	4.982759	1.959473	5	0.243017	Deposit and Collection Receptacles
54	4.982759	2.620392	4	0.216418	Harness
91	5.010574	3.974208	4	0.184555	Motors: Expansible Chamber Type
297	5.018194	2.242481	5	0.40425	Chairs and Seats
271	5.021898	2.575499	4	0.41299	Sheet Feeding or Delivering
108	5.028382	2.173272	5	0.316942	Horizontally Supported Planar Surfaces
238	5.033333	1.598889	6	0.078227	Railways: Surface Track
4	5.039892	2.738031	5	0.261231	Baths, Closets, Sinks, and Spittoons
190	5.043046	1.676955	5	0.367844	Trunks and Hand-Carried Luggage
475	5.073192	2.677623	5	0.485594	Planetary Gear Transmission Systems or Components
70	5.111992	2.556338	5	0.317345	Locks
182	5.124845	2.084537	5	0.207862	Fire Escape, Ladder, or Scaffold
142	5.142857	2.408163	6	0.116667	Wood Turning

Table S2 – A New Typology for Scientific Fields: Mean, mode, and standard deviation of the distance metric and percentage connectivity for all science and engineering WOS fields.

FieldCode	Dmean	DVariance	DMode	Connectivity	Field Name
'QE'	2.000455373	0.806668483	2	0.895838999	MATERIALS SCIENCE, BIOMATERIALS
'ES'	2.052515944	1.514040464	1	0.61741211	COMPUTER SCIENCE, HARDWARE & ARCHITECTURE
'ZE'	2.160138496	0.887671778	2	0.811426427	VIROLOGY
'YE'	2.163071314	1.573954645	1	0.619344606	TELECOMMUNICATIONS
'DX'	2.166021029	1.159640447	2	0.763447916	CHEMISTRY, CLINICAL & MEDICINAL
'RB'	2.187304426	1.274259922	2	0.444820044	ROBOTICS
'EW'	2.220658463	1.723829235	1	0.592035782	COMPUTER SCIENCE, SOFTWARE ENGINEERING
'IG'	2.229331439	1.073210106	2	0.755923766	ENGINEERING, BIOMEDICAL
'EP'	2.238099561	1.33062892	2	0.540551593	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE
'DR'	2.245792206	0.835648793	2	0.805939991	CELL BIOLOGY
'IQ'	2.246137132	1.531844372	2	0.664126678	ENGINEERING, ELECTRICAL & ELECTRONIC
'CQ'	2.254216773	0.799867238	2	0.833541259	BIOCHEMISTRY & MOLECULAR BIOLOGY
'NS'	2.258480208	1.198891536	2	0.96369104	NANOSCIENCE & NANOTECHNOLOGY
'YR'	2.265170407	1.627939019	1	0.298548207	TRANSPORTATION SCIENCE & TECHNOLOGY
'DB'	2.269017888	1.072221477	2	0.781278071	BIOTECHNOLOGY & APPLIED MICROBIOLOGY
'NI'	2.286904228	0.908191204	2	0.791996116	IMMUNOLOGY
'MA'	2.310863154	0.867490998	2	0.638326162	HEMATOLOGY
'CO'	2.317118557	1.064288758	2	0.762187289	BIOCHEMICAL RESEARCH METHODS
'DM'	2.357995621	0.896881859	2	0.785348848	ONCOLOGY
'DA'	2.359498136	0.739319422	2	0.826025086	BIOPHYSICS
'QU'	2.380705006	0.959156679	2	0.772255125	MICROBIOLOGY
'HQ'	2.384467829	1.017946848	2	0.819529285	ELECTROCHEMISTRY
'HY'	2.395286626	0.886467535	2	0.835591454	DEVELOPMENTAL BIOLOGY
'EE'	2.421917479	1.009018292	2	0.797615178	CHEMISTRY, ORGANIC
'QG'	2.432983193	1.129038159	2	0.743471198	MATERIALS SCIENCE, COATINGS & FILMS
'ZD'	2.457285654	1.025186866	2	0.630565881	PERIPHERAL VASCULAR DISEASE
'ET'	2.470588235	2.000632122	2	0.562881986	COMPUTER SCIENCE, INFORMATION SYSTEMS
'NN'	2.482106477	1.12244251	2	0.694167155	INFECTIOUS DISEASES
'QA'	2.489552493	1.047878997	2	0.714372115	MEDICINE, RESEARCH & EXPERIMENTAL
'EX'	2.492561634	1.915165282	2	0.456393664	COMPUTER SCIENCE, THEORY & METHODS
'IA'	2.498671521	0.831824298	2	0.76168954	ENDOCRINOLOGY & METABOLISM
'TU'	2.503186339	1.039774471	2	0.733145597	PHARMACOLOGY & PHARMACY
'UY'	2.504428246	1.065325203	2	0.788328086	POLYMER SCIENCE
'KM'	2.533144064	1.142282813	2	0.762421975	GENETICS & HEREDITY
'RU'	2.538575241	0.88800968	2	0.768934144	NEUROSCIENCES
'IK'	2.542159596	1.527101327	2	0.421904308	ENGINEERING, MANUFACTURING
'DS'	2.542421746	1.172888557	2	0.533475627	CRITICAL CARE MEDICINE
'WH'	2.544892389	1.077763171	2	0.578190249	RHEUMATOLOGY
'UB'	2.552574878	1.532896109	2	0.744703692	PHYSICS, APPLIED

Table S2 - Continued

FieldCode	Dmean	DVariance	DMode	Connectivity	Field Name
'PW'	2.552738199	0.99828973	2	0.637875603	MEDICAL LABORATORY TECHNOLOGY
'SY'	2.578029585	1.70677133	2	0.723990611	OPTICS
'DW'	2.599112377	1.230688733	2	0.69495076	CHEMISTRY, APPLIED
'ER'	2.60223719	1.65330748	2	0.58405339	COMPUTER SCIENCE, CYBERNETICS
'AQ'	2.630367727	1.034383232	2	0.622702621	ALLERGY
'UM'	2.64715362	0.839832768	2	0.777795452	PHYSIOLOGY
'DQ'	2.650212198	1.130483361	2	0.64738696	CARDIAC & CARDIOVASCULAR SYSTEMS
'WF'	2.653299372	1.044056464	2	0.731614301	REPRODUCTIVE BIOLOGY
'DY'	2.658871651	1.224748458	2	0.71431889	CHEMISTRY
'PK'	2.686910467	1.308532321	2	0.644788838	MATERIALS SCIENCE, CERAMICS
'WE'	2.688599161	1.085259143	2	0.649015164	RESPIRATORY SYSTEM
'OI'	2.707674944	1.42959497	2	0.426508344	INTEGRATIVE & COMPLEMENTARY MEDICINE
'KI'	2.712802855	1.113091668	2	0.547032657	GASTROENTEROLOGY & HEPATOLOGY
'YP'	2.718964313	1.09653419	3	0.685113169	TRANSPLANTATION
'MC'	2.724137931	1.295283393	3	0.939664804	MATHEMATICAL & COMPUTATIONAL BIOLOGY
'RO'	2.736249038	1.825403492	2	0.659792824	MULTIDISCIPLINARY SCIENCES
'JY'	2.740697828	1.250923136	3	0.717134501	FOOD SCIENCE & TECHNOLOGY
'SU'	2.743488889	1.166304537	2	0.66994619	OPHTHALMOLOGY
'RT'	2.746485338	1.117390192	3	0.618586666	CLINICAL NEUROLOGY
'TM'	2.748943517	1.069972585	3	0.715021032	PATHOLOGY
'EA'	2.752290251	1.287565739	3	0.714480464	CHEMISTRY, ANALYTICAL
'II'	2.770698448	1.361841161	2	0.649307074	ENGINEERING, CHEMICAL
'EI'	2.777184388	1.18233288	3	0.773629456	CHEMISTRY, PHYSICAL
'SA'	2.777351401	1.183871008	3	0.724844623	NUTRITION & DIETETICS
'AA'	2.777882593	1.526160585	2	0.622163259	ACOUSTICS
'TC'	2.783639515	1.313867516	2	0.702165151	ORTHOPEDICS
'RX'	2.787536015	1.167643176	3	0.656462347	NEUROIMAGING
'GA'	2.791005838	1.282172553	3	0.560909881	DERMATOLOGY
'XW'	2.792349271	1.097937585	3	0.637940675	SPORT SCIENCES
'PM'	2.79539511	1.520616949	2	0.692062033	MATERIALS SCIENCE
'RA'	2.813261824	1.115191154	3	0.716381418	MICROSCOPY
'AE'	2.813704994	1.756699513	2	0.53010713	AGRICULTURAL ENGINEERING
'AZ'	2.814777498	1.005994892	3	0.759372609	ANDROLOGY
'AC'	2.823231094	1.759039998	2	0.491995688	AUTOMATION & CONTROL SYSTEMS
'ZA'	2.827878268	1.287366561	3	0.64333433	UROLOGY & NEPHROLOGY
'UE'	2.833733014	1.815520757	3	0.6502079	IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY
'OA'	2.837660851	1.669107831	3	0.693409797	INSTRUMENTS & INSTRUMENTATION
'VY'	2.844080131	1.37279136	3	0.662555289	RADIOLOGY & NUCLEAR MEDICINE
'ID'	2.850692958	1.672591695	3	0.641989137	ENERGY & FUELS
'BA'	2.854042061	1.234436199	3	0.555049851	ANESTHESIOLOGY
'AH'	2.883350104	1.873148743	2	0.503841504	AGRICULTURE, MULTIDISCIPLINARY
'YO'	2.88610166	1.088541528	3	0.746718631	TOXICOLOGY

Table S2 - Continued

FieldCode	Dmean	DVariance	DMode	Connectivity	Field Name
'FF'	2.897294589	1.375795715	3	0.602618016	CRITICAL CARE
'PT'	2.905431272	1.797792011	3	0.658785514	MEDICAL INFORMATICS
'IP'	2.931407233	2.024776164	3	0.522301494	ENGINEERING, PETROLEUM
'EC'	2.936963046	1.228712323	3	0.728637023	CHEMISTRY, INORGANIC & NUCLEAR
'EV'	2.947590967	1.86971959	3	0.587870073	COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
'FQ'	2.94922835	1.111990214	3	0.800645364	CYTOLOGY & HISTOLOGY
'AY'	2.957326704	0.99136887	3	0.671284547	ANATOMY & MORPHOLOGY
'YA'	2.961602875	1.284787603	3	0.679734879	SURGERY
'DE'	2.964523994	1.400225748	3	0.727387685	PLANT SCIENCES
'IH'	2.972563077	1.535135813	3	0.62832441	ENGINEERING, ENVIRONMENTAL
'CU'	2.980614199	1.42057425	3	0.570669162	BIOLOGY
'QH'	2.98740993	1.348879419	3	0.525376487	MATERIALS SCIENCE, COMPOSITES
'SD'	2.99151309	1.216292759	3	0.687575458	OBSTETRICS & GYNECOLOGY
'SR'	2.998209223	1.681238399	3	0.684683764	REMOTE SENSING
'FY'	3.010897544	1.328972634	3	0.643507876	DENTISTRY, ORAL SURGERY & MEDICINE
'FI'	3.01364222	1.408117222	3	0.671993462	CRYSTALLOGRAPHY
'QJ'	3.013841567	1.652704493	3	0.579145341	MATERIALS SCIENCE, TEXTILES
'XQ'	3.017056183	1.451073684	3	0.701829979	SPECTROSCOPY
'QF'	3.041162608	1.799719622	3	0.47358631	MATERIALS SCIENCE, CHARACTERIZATION & TESTING
'PJ'	3.048099484	1.841891038	3	0.637665981	MATERIALS SCIENCE, PAPER & WOOD
'UH'	3.099336432	1.125456617	3	0.75673313	PHYSICS, ATOMIC, MOLECULAR & CHEMICAL
'PE'	3.112900438	1.772947199	3	0.519741866	OPERATIONS RESEARCH & MANAGEMENT SCIENCE
'PY'	3.126473867	1.48725224	3	0.500184167	MEDICINE, GENERAL & INTERNAL
'IF'	3.127803179	1.98571294	3	0.593257556	ENGINEERING
'UK'	3.131614517	1.564238525	3	0.699758767	PHYSICS, CONDENSED MATTER
'CX'	3.135644632	1.258484549	3	0.737305966	BIOLOGY, MISCELLANEOUS
'TD'	3.138106711	1.314560275	3	0.665544593	OTORHINOLARYNGOLOGY
'IU'	3.14595263	1.844639272	3	0.508427261	ENGINEERING, MECHANICAL
'TQ'	3.151048542	1.264168719	3	0.577259247	PEDIATRICS
'IJ'	3.161330475	1.796460549	3	0.419439046	ENGINEERING, INDUSTRIAL
'JI'	3.170657299	1.604970567	3	0.641615754	ERGONOMICS
'LI'	3.170760516	1.511712232	3	0.689577934	GERIATRICS & GERONTOLOGY
'EY'	3.173393461	1.688311258	2	0.617258177	COMPUTER APPLICATIONS & CYBERNETICS
'ZC'	3.181605051	1.659014304	3	0.595904779	VETERINARY SCIENCES
'IO'	3.182297155	2.130097568	3	0.356901091	ENGINEERING, OCEAN
'TI'	3.183465784	1.600578717	3	0.708565351	PARASITOLOGY
'CN'	3.198730773	1.230208615	3	0.708936398	BEHAVIORAL SCIENCES
'IE'	3.208333333	1.081597222	3	1.043478261	*ENGINEERING & TECHNOLOGY
'DT'	3.21080245	1.666213143	3	0.526152864	THERMODYNAMICS
'BV'	3.215109859	1.166362454	3	0.708378291	PSYCHOLOGY, BIOLOGICAL
'YU'	3.221126992	1.24678217	3	0.69260178	TROPICAL MEDICINE

Table S2 - Continued

FieldCode	Dmean	DVariance	DMode	Connectivity	Field Name
'PZ'	3.224847978	1.743789045	3	0.561281765	METALLURGY & METALLURGICAL ENGINEERING
'UF'	3.225013679	1.401995144	3	0.685860818	PHYSICS, FLUIDS & PLASMAS
'AD'	3.241166449	1.503276851	3	0.63711656	AGRICULTURE, DAIRY & ANIMAL SCIENCE
'QM'	3.248434985	1.27224536	3	0.765848447	METALLURGY & MINING
'AI'	3.250647788	2.055859751	3	0.472527713	AEROSPACE ENGINEERING & TECHNOLOGY
'JA'	3.281493464	1.542458927	3	0.666528701	ENVIRONMENTAL SCIENCES
'RQ'	3.283547973	1.520760161	3	0.651259894	MYCOLOGY
'PU'	3.288005342	1.63562281	3	0.579782228	MECHANICS
'NE'	3.342865912	1.45110147	3	0.640182803	PUBLIC HEALTH
'ZR'	3.357453299	1.686850005	3	0.6106582	WATER RESOURCES
'IL'	3.360163383	1.946928181	3	0.600230747	ENGINEERING, MARINE
'FA'	3.36234502	2.054189295	3	0.435586552	CONSTRUCTION & BUILDING TECHNOLOGY
'IX'	3.364720395	1.936469165	3	0.230521327	ENGINEERING, GEOLOGICAL
'GC'	3.365216217	1.595969243	3	0.665093786	GEOCHEMISTRY & GEOPHYSICS
'QQ'	3.366254571	1.451860928	3	0.737950659	METEOROLOGY & ATMOSPHERIC SCIENCES
'GM'	3.377040089	1.363624324	3	0.657669862	SUBSTANCE ABUSE
'RY'	3.382648589	1.916142152	3	0.624800691	NUCLEAR SCIENCE & TECHNOLOGY
'YQ'	3.390052833	2.504626869	3	0.545030285	TRANSPORTATION
'LJ'	3.400032965	1.694102521	3	0.417406261	GERONTOLOGY
'OU'	3.407975742	1.47244213	3	0.673098927	LIMNOLOGY
'ZQ'	3.409793046	1.576150294	3	0.608094875	MINING & MINERAL PROCESSING
'XY'	3.431073209	1.845658148	3	0.566643557	STATISTICS & PROBABILITY
'IM'	3.43504973	1.869205234	3	0.500790705	ENGINEERING, CIVIL
'RE'	3.480622747	1.586673532	3	0.665718455	MINERALOGY
'SI'	3.489465639	1.520819864	3	0.682039487	OCEANOGRAPHY
'XE'	3.496908409	1.353429089	3	0.699939968	AGRICULTURE, SOIL SCIENCE
'HT'	3.52486106	1.630634243	3	0.572501273	EVOLUTIONARY BIOLOGY
'IY'	3.530509878	1.693825204	3	0.70478471	ENTOMOLOGY
'PO'	3.534453205	1.817033751	3	0.563780441	MATHEMATICS, INTERDISCIPLINARY APPLICATIONS
'PC'	3.603909142	1.948244742	3	0.514769295	MANAGEMENT
'OP'	3.606965899	2.198773977	3	0.59270136	MEDICINE, LEGAL
'LE'	3.61926219	1.604044105	3	0.627690397	GEOSCIENCES, INTERDISCIPLINARY
'QB'	3.620943953	1.766935982	3	0.640468543	MEDICINE, MISCELLANEOUS
'UI'	3.625404929	1.818091923	3	0.647920952	PHYSICS
'PI'	3.626846826	1.462690223	3	0.69587568	MARINE & FRESHWATER BIOLOGY
'HL'	3.715426825	2.097874814	3	0.412108201	HEALTH CARE SCIENCES & SERVICES
'UR'	3.762748455	2.04934708	3	0.616860454	PHYSICS, MATHEMATICAL
'PN'	3.768383547	2.473276646	3	0.453640155	MATHEMATICS, APPLIED
'KY'	3.86020054	1.564605745	4	0.614222096	GEOLOGY
'PS'	3.889082735	1.825917853	4	0.549075234	SOCIAL SCIENCES, MATHEMATICAL METHODS
'VS'	3.913568851	1.594946372	4	0.724261578	PSYCHOLOGY, MATHEMATICAL
'LQ'	3.922052935	1.865537796	3	0.542757959	HEALTH POLICY & SERVICES

Table S2 - Continued

FieldCode	Dmean	DVariance	DMode	Connectivity	Field Name
'BU'	3.942043558	1.424864062	4	0.721928768	ASTRONOMY & ASTROPHYSICS
'UN'	3.955609085	1.594027278	4	0.685529241	PHYSICS, NUCLEAR
'OY'	4.063573523	2.224487942	3	0.489904821	LANGUAGE & LINGUISTICS
'WV'	4.084229246	1.465704723	4	0.548287019	SOCIAL SCIENCES, BIOMEDICAL
'AK'	4.106824926	1.644374785	4	0.846733668	AGRICULTURAL EXPERIMENT STATION REPORTS
'OO'	4.117370892	1.969009676	4	0.193519079	MEDICAL ETHICS
'UP'	4.132496735	1.925878701	4	0.673321896	PHYSICS, PARTICLES & FIELDS
'NU'	4.141747868	3.43072971	4	0.37384738	INFORMATION SCIENCE & LIBRARY SCIENCE
'ZI'	4.146067416	2.326978917	3	0.74789916	WELDING TECHNOLOGY
'BD'	4.241275996	2.159445393	4	0.369304869	BIODIVERSITY CONSERVATION
'KV'	4.29750361	2.644728656	4	0.364902507	GEOGRAPHY, PHYSICAL
'RZ'	4.332382636	2.043591716	4	0.451352952	NURSING
'NQ'	4.430458289	2.023893222	4	0.56823852	PSYCHOLOGY, APPLIED
'KU'	4.596694694	2.324249247	4	0.525188437	GEOGRAPHY
'EU'	4.609385113	2.134636839	4	0.451160753	COMMUNICATION
'AF'	4.707584393	2.270038926	4	0.457343358	AGRICULTURAL ECONOMICS & POLICY
'JB'	4.726622381	2.032060794	4	0.428755306	ENVIRONMENTAL STUDIES
'PQ'	4.916812289	2.991413612	4	0.375614875	MATHEMATICS

Table S3A – Patent Home Run Regressions

VARIABLES	(1) Home Run	(2) Home Run	(3) Home Run	(4) Home Run	(5) Home Run	(6) Home Run	(7) Home Run
D = 1	0.0255***	0.0313***	0.0249***	0.0210***	0.0235***	0.0236***	0.0180***
	(0.000517)	(0.000562)	(0.000517)	(0.000516)	(0.000519)	(0.000523)	(0.000738)
Disconnected	-0.0107***	-0.0174***	-0.0229***	-0.00186***	-0.00851***	-0.0112***	-0.0196***
	(0.000279)	(0.000317)	(0.000343)	(0.000336)	(0.000281)	(0.000279)	(0.000545)
Constant	0.0546***	0.0562***	0.0587***	0.0521***	0.0541***	0.0549***	0.0584***
	(0.000182)	(0.000192)	(0.000201)	(0.000190)	(0.000181)	(0.000183)	(0.000249)
Class	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Inventors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	2,813,208	2,813,196	2,813,208	2,813,208	2,813,208	2,813,208	2,813,196
R-squared	0.002	0.204	0.044	0.082	0.003	0.041	0.445

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S3B – Patent Impact Regressions using Alternative Impact Measures (log citations)

VARIABLES	(1) lognCite8	(2) lognCite8	(3) lognCite8	(4) lognCite8	(5) lognCite8	(6) lognCite8	(7) lognCite8
D = 1	0.0813***	0.146***	0.0216***	0.0690***	0.0769***	0.0536***	0.100***
	(0.00224)	(0.00235)	(0.00185)	(0.00202)	(0.00225)	(0.00222)	(0.00265)
Disconnected	-0.00973***	0.144***	-0.402***	-0.222***	-0.00434***	-0.0202***	-0.143***
	(0.00121)	(0.00129)	(0.00137)	(0.00140)	(0.00123)	(0.00121)	(0.00216)
Constant	1.233***	1.174***	1.371***	1.306***	1.232***	1.240***	1.276***
	(0.000850)	(0.000833)	(0.000761)	(0.000807)	(0.000850)	(0.000832)	(0.000943)
Class	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Inventors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	2,813,208	2,813,196	2,813,208	2,813,208	2,813,208	2,813,208	2,813,196
R-squared	0.001	0.046	0.288	0.187	0.001	0.032	0.606

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S3C – Patent Impact Regressions using Alternative Impact Measures (Home Run at 1% threshold)

VARIABLES	(1) Home Run (1%)	(2) Home Run (1%)	(3) Home Run (1%)	(4) Home Run (1%)	(5) Home Run (1%)	(6) Home Run (1%)	(7) Home Run (1%)
D = 1	0.00963***	0.0105***	0.00938***	0.00775***	0.00909***	0.00939***	0.00650***
	(0.000270)	(0.000302)	(0.000270)	(0.000267)	(0.000271)	(0.000275)	(0.000376)
Disconnected	-0.00734***	-0.00136***	-0.0122***	-0.00560***	-0.00675***	-0.00753***	-0.00249***
	(0.000107)	(0.000115)	(0.000147)	(0.000130)	(0.000108)	(0.000108)	(0.000190)
Constant	0.0114***	0.00929***	0.0131***	0.0110***	0.0113***	0.0115***	0.0101***
	(8.51e-05)	(7.86e-05)	(9.75e-05)	(8.80e-05)	(8.45e-05)	(8.56e-05)	(9.78e-05)
Class	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Inventors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	2,813,208	2,813,196	2,813,208	2,813,208	2,813,208	2,813,208	2,813,196
R-squared	0.003	0.121	0.041	0.077	0.003	0.045	0.488

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S4A – Paper Home Run Regressions

VARIABLES	(1) Home Run	(2) Home Run	(3) Home Run	(4) Home Run	(5) Home Run	(6) Home Run	(7) Home Run
D = 1	0.131***	0.137***	0.129***	0.126***	0.127***	0.131***	0.1238***
	(0.000353)	(0.000356)	(0.000355)	(0.000350)	(0.000352)	(0.000353)	(0.000370)
Disconnected	-0.0262***	-0.0300***	-0.0509***	-0.0326***	-0.0287***	-0.0257***	-0.0454***
	(8.53e-05)	(9.09e-05)	(0.000111)	(8.58e-05)	(8.55e-05)	(8.47e-05)	(0.000103)
Constant	0.0568***	0.0580***	0.0673***	0.0597***	0.0581***	0.0565***	0.0531***
	(6.56e-05)	(6.76e-05)	(7.79e-05)	(6.70e-05)	(6.63e-05)	(6.53e-05)	(5.35e-05)
Field	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144
R-squared	0.024	0.026	0.028	0.040	0.029	0.024	0.366

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S4B – Paper Impact Regressions using Alternative Impact Measures (log citations)

VARIABLES	(1) lognCite8	(2) lognCite8	(3) lognCite8	(4) lognCite8	(5) lognCite8	(6) lognCite8	(7) lognCite8
D = 1	0.726***	0.698***	0.593***	0.641***	0.624***	0.707***	0.503***
	(0.00110)	(0.00104)	(0.00111)	(0.00102)	(0.00107)	(0.00110)	(0.00117)
Disconnected	-0.674***	-0.580***	-1.024***	-0.749***	-0.717***	-0.634***	-0.742***
	(0.000393)	(0.000395)	(0.000462)	(0.000367)	(0.000383)	(0.000394)	(0.000592)
Constant	2.348***	2.309***	2.502***	2.384***	2.371***	2.332***	2.388***
	(0.000279)	(0.000268)	(0.000294)	(0.000259)	(0.000272)	(0.000277)	(0.000316)
Field	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144
R-squared	0.150	0.216	0.239	0.265	0.201	0.169	0.552

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S4C – Paper Impact Regressions using Alternative Impact Measures (Home Run at 1% threshold)

VARIABLES	(1) Home Run (1%)	(2) Home Run (1%)	(3) Home Run (1%)	(4) Home Run (1%)	(5) Home Run (1%)	(6) Home Run (1%)	(7) Home Run (1%)
D = 1	0.0390***	0.0390***	0.0396***	0.0380***	0.0378***	0.0391***	0.0390***
	(0.000193)	(0.000193)	(0.000196)	(0.000193)	(0.000193)	(0.000194)	(0.000193)
Disconnected	-0.00409***	-0.00258***	-0.00620***	-0.00522***	-0.00481***	-0.00413***	-0.00409***
	(3.69e-05)	(3.85e-05)	(5.52e-05)	(3.79e-05)	(3.70e-05)	(3.70e-05)	(3.69e-05)
Constant	0.00986***	0.00922***	0.0107***	0.0104***	0.0102***	0.00987***	0.00986***
	(2.80e-05)	(2.77e-05)	(3.47e-05)	(2.89e-05)	(2.85e-05)	(2.80e-05)	(2.80e-05)
Field	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144	23,690,144
R-squared	0.009	0.016	0.009	0.012	0.012	0.009	0.344

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S5 – Patent Maintenance Fee Regressions

VARIABLES	(1) No. of Times Maintenance Fee Paid	(2) No. of Times Maintenance Fee Paid	(3) No. of Times Maintenance Fee Paid	(4) No. of Times Maintenance Fee Paid	(5) No. of Times Maintenance Fee Paid	(6) No. of Times Maintenance Fee Paid
D = 1	0.0570***	0.0807***	0.0874***	0.0374***	0.0323***	0.0412***
	(0.00221)	(0.00235)	(0.00177)	(0.00211)	(0.00221)	(0.00280)
Disconnected	-1.027***	-0.955***	-0.219***	-0.673***	-1.000***	-0.0482***
	(0.00132)	(0.00146)	(0.00129)	(0.00158)	(0.00134)	(0.00218)
Constant	1.464***	1.440***	1.225***	1.363***	1.459***	1.181***
	(0.000913)	(0.000920)	(0.000704)	(0.000913)	(0.000911)	(0.000912)
Class	No	Yes	No	No	No	Yes
Year	No	No	Yes	No	No	Yes
No. of Refs.	No	No	No	Yes	No	Yes
No. of Inventors	No	No	No	No	Yes	Yes
Observations	2,615,177	2,615,171	2,615,177	2,615,177	2,615,177	2,615,171
R-squared	0.167	0.195	0.470	0.249	0.172	0.679

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Notes: The dependent variable is the number of times the patent maintenance fees are paid (with renewal fees due in the 4th, 8th, and 12th year after the patent was granted for U.S. patents).

Table S6 – Patent Impact Regressions with Individual Inventor Fixed Effects

VARIABLES	(1) Home Run (5%)	(2) Home Run (5%)	(3) Home Run (5%)	(4) Home Run (5%)
D = 1	0.0183***	0.0171***	0.0191***	0.0163***
	(0.000689)	(0.000928)	(0.00104)	(0.00154)
Constant	0.0786***	0.0791***	0.0783***	0.0794***
	(0.000377)	(0.000440)	(0.000476)	(0.000641)
Individual	Yes	Yes	Yes	Yes
Field	No	Yes	No	Yes
Year	No	No	Yes	Yes
Observations	963,905	963,904	963,905	963,904
R-squared	0.284	0.601	0.664	0.858

Notes: Regression sample considers connected patents, with individual fixed effects for the individual inventor. Observations are at the individual by patent level. Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table S7 – Paper Impact Regressions with Individual Author Fixed Effects

VARIABLES	(1) Home Run (5%)	(2) Home Run (5%)	(3) Home Run (5%)	(4) Home Run (5%)
D = 1	0.139***	0.146***	0.141***	0.148***
	(0.000141)	(0.000163)	(0.000155)	(0.000219)
Constant	0.0624***	0.0615***	0.0622***	0.0614***
	(4.72e-05)	(4.75e-05)	(4.78e-05)	(5.09e-05)
Individual	Yes	Yes	Yes	Yes
Field	No	Yes	No	Yes
Year	No	No	Yes	Yes
Observations	35,502,198	35,502,198	35,502,198	35,502,198
R-squared	0.108	0.313	0.250	0.638

Notes: Regression sample considers connected papers, with individual fixed effects for the individual authors. Observations are at the individual by paper level. Robust standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table S8 – Institutional Type and Distance: Papers

VARIABLES	(1) D	(2) D	(3) D	(4) D	(5) D	(6) D	(7) D
Univ	0.358***	0.229***	0.348***	0.415***	0.297***	0.396***	0.113***
	(0.00173)	(0.00157)	(0.00171)	(0.00172)	(0.00170)	(0.00158)	(0.00286)
Gov	0.430***	0.134***	0.430***	0.482***	0.400***	0.526***	0.101***
	(0.00334)	(0.00308)	(0.00332)	(0.00333)	(0.00331)	(0.00308)	(0.00487)
Constant	2.527***	2.652***	2.537***	2.474***	2.584***	2.491***	2.760***
	(0.00168)	(0.00152)	(0.00165)	(0.00167)	(0.00165)	(0.00152)	(0.00269)
Field	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
No. of Citations	No	No	No	No	No	Yes	Yes
Observations	9,354,919	9,354,919	9,354,919	9,354,919	9,354,919	9,354,919	9,354,919
R-squared	0.005	0.242	0.023	0.028	0.069	0.198	0.866

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S9 – Institutional Type and Distance: Patents

VARIABLES	(1) D	(2) D	(3) D	(4) D	(5) D	(6) D	(7) D
Univ	-0.847***	-0.468***	-0.846***	-0.866***	-0.838***	-0.861***	-0.442***
	(0.00480)	(0.00481)	(0.00481)	(0.00483)	(0.00481)	(0.00485)	(0.0231)
Gov	-0.411***	-0.271***	-0.397***	-0.472***	-0.422***	-0.447***	-0.360***
	(0.0100)	(0.00912)	(0.0100)	(0.0100)	(0.00996)	(0.0100)	(0.0472)
Constant	2.583***	2.536***	2.582***	2.586***	2.582***	2.585***	2.535***
	(0.00199)	(0.00170)	(0.00198)	(0.00198)	(0.00199)	(0.00199)	(0.00352)
Class	No	Yes	No	No	No	No	Yes
Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Inventors	No	No	No	No	Yes	No	Yes
No. of Citations	No	No	No	No	No	Yes	Yes
Observations	1500,943	1500,942	1500,943	1500,943	1500,943	1500,943	1500,942
R-squared	0.044	0.298	0.053	0.066	0.047	0.052	0.933

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S10A – Home Run Rate for $D = 1$ Same Author-Inventor Paper vs. other $D = 1$ Papers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	homerun	homerun	homerun	homerun	homerun	homerun	homerun
Same Individual	0.0531***	0.0555***	0.0560***	0.0572***	0.0518***	0.0535***	0.0581***
	(0.00134)	(0.00134)	(0.00135)	(0.00132)	(0.00133)	(0.00134)	(0.00271)
Constant	0.184***	0.183***	0.183***	0.183***	0.184***	0.184***	0.183***
	(0.000359)	(0.000357)	(0.000359)	(0.000355)	(0.000357)	(0.000359)	(0.000387)
Field	No	Yes	No	No	No	No	Yes
Pub Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	1,269,687	1,269,687	1,269,687	1,269,687	1,269,687	1,269,687	1,269,687
R-squared	0.001	0.013	0.004	0.024	0.013	0.002	0.674

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S10B – Home Run Rate for $D = 1$ Same Author-Inventor Patent vs. other $D = 1$ Patents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	homerun	homerun	homerun	homerun	homerun	homerun	homerun
Same Individual	0.00225**	0.0101***	0.00241**	0.00432**	0.000624	0.00273**	-0.00124
	(0.00114)	(0.00120)	(0.00114)	(0.00113)	(0.00115)	(0.00118)	(0.00242)
Constant	0.0794***	0.0775***	0.0794***	0.0789***	0.0798***	0.0793***	0.0803***
	(0.000554)	(0.000551)	(0.000554)	(0.000548)	(0.000556)	(0.000557)	(0.000712)
Field	No	Yes	No	No	No	No	Yes
Pub Year	No	No	Yes	No	No	No	Yes
No. of Refs	No	No	No	Yes	No	No	Yes
No. of Authors	No	No	No	No	Yes	No	Yes
Institution	No	No	No	No	No	Yes	Yes
Observations	313,921	313,921	313,921	313,921	313,921	313,921	313,921
R-squared	0.001	0.008	0.001	0.019	0.001	0.002	0.764

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1