Reliance on Science in Patenting:

USPTO Front-Page Citations to Scientific Articles

Matt Marx† & Aaron Fuegi††

27 August 2019

Abstract: To what extent do firms rely on basic science in their R&D efforts? Several scholars have sought to answer this and related questions, but progress has been impeded by the difficulty of matching unstructured references in patents to published papers. We introduce an open-access dataset of references from the front pages of all patents granted by the U.S. Patent & Trademark Office to scientific papers published since 1800 as captured by the Microsoft Academic Graph. Each patent-paper linkage is assigned a confidence score, which is characterized in a random sample by false negatives vs. false positives. We outline several avenues for strategy research enabled by these new data.

† Boston University Questrom School of Business; †† Boston University IS&T Research Computing Services group; feedback to mattmarx@bu.edu. We thank Kysha Johnson, Erin Thomas, and especially Dmitrii Shelekhov for assistance in constructing the list of known-good references. We are grateful to Guan-Cheng Li for sharing the unstructured non-patent references extracted from raw USPTO data. The authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster which is administered by Boston University's Research Computing Services; in particular, we thank Katia Oleinik, Charles Jahnke, and Wayne Gilmore of IS&T Research Computing Services for distributed computing support. Errors are ours.

INTRODUCTION

This paper details the construction of a publicly-available set of citations from U.S. patents (1947-2018) to scientific articles (1800-2018). We establish approximately 16.7MM patent citations to science. The patent-paper linkages, as well as selected metadata on the articles (whether cited or not), and the source code are publicly available for download at **http://relianceonscience.org**.

Patent citations to science (hereafter, PCS) are of interest to strategy researchers who seek to understand innovation in firms: the nature of research and development, how inventors and scientists search for commercializable basic science, and the process by which university inventions are exploited by firms. Despite these advantages, PCS have only sometimes been used in strategy research, for at least two reasons. First, PCS are difficult to work with given that they appear in patent records as unstructured text strings. Thus researchers must either match patents and scientific articles by hand (for small samples) or (for large samples) build algorithms that are possibly error prone. Second, even when research teams have invested the effort to link patents and scientific articles at scale, they have typically done so using proprietary databases such as Scopus or the Web of Science. Thus the matched PCS cannot be shared with other research teams, who must license the databases for themselves and/or develop algorithms from scratch.

As other research teams have (Gaetani and Bergolis, 2015; Fleming et al., 2018), we link data from the U.S. Patent & Trademark Office to a broad set of scientific articles not limited by industry or field. Specifically, we cover all U.S. patents from 1947-2018, correcting for many errors in OCRed data prior to 1976. Our linkages involve not only proprietary article databases, which cannot be shared, but also a newly-available, open-source database from Microsoft (Sinha et al, 2015) which permits us to post the resulting PCS for public use. Based on third-party

assessment, we estimate that our algorithm can capture up to 93% of patent citations to science with an accuracy rate of 99% or higher. We believe this to be the longest panel of patent-to-paper citations (spanning more than seven decades) that is publicly available and is accompanied by rigorous performance metrics.

The paper is organized as follows. We begin by motivating the use of PCS in strategy research and review prior approaches. Second, we detail our patent-paper linking algorithm. Third, we describe both the private and publicly-available data products as well as our methods for assessing their efficacy. We conclude by sketching research avenues opened up by the broad availability of PCS.

MOTIVATION

Innovation is a key source of sustainable differentiation for firms and thus a longtime focus of strategy researchers. The lottery-based nature of research & development (R&D) has long prompted inquiry into the nature of the inventive process (Scherer, 2001), including both how internal R&D projects are managed and how external sources of commercializable science are accessed (Nelson, 1982; Mokyr, 2002; Cohen, Nelson, & Walsh, 2002; Fleming & Sorenson, 2004).

Given that firms appear to be retreating from investing internally in basic science (Arora, Belenzon, & Pattaconi, 2018), it is perhaps more important than ever to understand to where firms look for technological inspiration, as well as how they differentiate themselves from that source material in order to secure temporary monopoly rights in the form of intellectual property protection. The growth of markets for technology (Arora, Cohen, & Walsh, 2016; Arora, Fosfuri, & Gambardella, 2001), including from academia, thanks to changing norms and policies

including the Bayh-Dole Act, entail that firms can sample from a larger scientific palette than ever, including established firms, startups, government agencies (Fleming, et al., 2019), and "lone" inventors. Firms are moreover thought to configure themselves to more easily engage with external innovators and absorb their knowledge (Gambardella, 1992; Cockburn & Henderson, 1998; Cohen & Levinthal, 1991).

But tracing the scientific lineage of R&D—whether inside or outside the firm—can be elusive. Firms are under no obligation to disclose where their innovations came from, except in the case of patented inventions. Of course not every innovation is patented, and many questionable patents are granted. But the process of prosecuting a patent—especially at the U.S. Patent and Trademark Office—obligates the applicant to disclose "prior art" against which the focal invention is distinguished and upon which the inventors may have relied in their own inventive process. Applicants to the USPTO are obligated "to disclose to the Office all information known to that individual to be material to patentability" (see http://www.uspto.web/offices/pac/mpep/mpep-2000.pdf). Prior art exists in two primary forms: a) references to prior patents b) references to non-patent literature (NPL). Because the omission of prior art, whether patent or non-patent can threaten the legitimacy of the patent, applicants have strong incentives to list all relevant patents as well as non-patent literature.

References to patents may provide clues to underlying technologies that influence a firm's own patented inventions, but by definition they provide a rather incomplete record. Belenzon and Schankerman (2013) suggest that barely 10% of scientific discoveries at universities are patented. Patenting inventors may have built upon a much wider array of basic science and technology than is captured by the patent corpus. Indeed, Roach and Cohen conduct a survey of R&D managers, finding that "citations to nonpatent references, such as scientific journal articles,

correspond more closely to managers' reports of the use of public research than do the more commonly employed citations to patent references" (Roach and Cohen 2013:505).

Hence, citations from patents to scientific articles can help expand our understanding of inputs into the R&D process, at least as far as the applicants were aware enough of prior art to report it in patent applications. Indeed, scholars have found PCS useful in furthering at least three research agendas; 1) describing the process of searching for innovations 2) characterizing the nature of R&D portfolios 3) the localization of spillovers from academia to industry.

Prior strategy work using PCS

Regarding search processes, Fleming and Sorenson (2004) use counts of NPLs from May and June 1990 to argue that science can serve as a "map" to help commercial inventors navigate the complexities of interdependent technologies. Counts of NPLs are also employed by Arts and Fleming (2018) to show that the negative effect of exploration on breakthroughs is mitigated by reliance on science. Katila and Ahuja (2002) dig deeper into R&D search processes by mapping NPL references to the original scientific papers for 124 robotics firms, contrasting deep search vs. search of wider scope. Gittelman & Kogut (2003) likewise trace NPL for patents from 116 biotech firms to describe selection logics of inventors, whose reliance on important scientific papers is negatively correlated with high-impact inventions.

PCS have also been used to characterize R&D more generally. Veugelers, Wang, and Stephan (2017) use combinations of papers cited by patents to measure novelty. Bransetter and Kwon (2004) show that the connection between patenting and science among 300 Japanese firms has contributed to both productivity and an increase in alliances. McMillan, Narin, & Deeds (2000) trace NPLs from 199 biotech firms that completed an IPO to show that these firms relied on very

basic research as compared with more applied work. Ribiero, et al (2014) further characterize the reliance of R&D in multinational corporations on cross-national networks using NPLs. Arora, Belenzon, and Sheer (2017) collect PCS for 4,736 firms from 1980-2006 to demonstrate that firms whose patents cite their own scientific papers invest more in R&D generally.

A third area of research supported by PCS is of spillovers from academia and government to industry, including the localization of such. Especially as firms retreat from investing internally in R&D, government and academia becoming primary sources of material upon which commercial inventors can build. Belenzon and Schankerman (2013) use both patent-to-patent citations and PCS to papers from 184 research universities to establish that the flow of university knowledge is geographically bounded. Li, Azoulay, & Sampat (2017) show that about 10% of NIH grants lead to a patent (as tracked via PCS). Fleming et al. (2019) document that nearly one-third of U.S. patents depend in some way on federally-funded research, including by the inclusion of a citation to a paper with a government grant. Ahmadpoor & Jones (2017) calculate the citation distance from papers to patents using PCS in order to show how deeply various fields rely on science.

The foregoing makes clear that strategy researchers find PCS useful. Indeed, the number of papers that have relied on PCS might call into question the need for the present exercise. Although many of the aforementioned papers are highly cited, it is difficult for researchers to replicate or build directly on them because the PCS used are either a) limited to counts of references b) limited to a small number of firms c) unavailable due to licensing restrictions.

Some papers use simple counts of NPL references from a patent as evidence of reliance on science, although many NPL references do not refer to scientific documents but product brochures, trade magazines, websites, and other non-patented material. Callaert, Grouwels, and

Van Looy (2011) have distinguished *scientific* NPLs from non-scientific ones, which represents a step forward, but several papers still use counts of such NPLs without linking to the scientific papers themselves that are referenced.

Some researchers have undertaken the task of mapping NPLs to scientific papers (e.g., Katila & Ahuja (2002), Branstetter & Kwon (2004), Gittelman & Kogut (2004), among others), but typically this has been undertaken for a limited number of firms in a single industry. Tjissen (2001) does so for a somewhat larger sample of Dutch research papers from 1987-1996. Hu et al. (2007) generate linkages to papers from 50,000 nanoscale engineering patents. Belenzon, and Sheer (2017) assemble PCS for 4,736 firms from 1980-2006.

In a few cases, researchers have mapped a comprehensive set of NPLs to scientific references, including Narin & Olivastro (1998), Gaetani & Bergolis (2015), Shirable (2014), Patelli et al (2017), Knaus & Palzenberger (2018), and Fleming et al., (2019). However, these linkages have been made to proprietary datasets such as the Clarivate Web of Science or Scopus and cannot be shared publicly. This presents two barriers for researchers who wish to verify or extend prior findings using PCS. First, they must pay to license the proprietary databases, which can be prohibitively expensive. Second, they must either obtain the code for the patent-paper linking algorithm from its developers or invest in creating their own linking algorithm.

Linking PCS to open datasets including the Microsoft Academic Graph

An alternative is to link PCS to open datasets. At least two research teams have linked to PubMed, which covers more than 20 million papers in the life sciences and can be downloaded at https://www.nlm.nih.gov/databases/download/pubmed_medline.html. Azoulay, Graff Zivin, and Sampat (2011) as well as Agarwal, et al (2011) have linked to various editions of PubMed.

Although it would be possible for other researchers to build directly on this work, to our knowledge neither the patent-to-PubMed linkages nor the code for generating these appears to be publicly available from either effort.

One might consider linking PCS to Google Scholar, a well-known repository of academic publications. However, Google obstructs users from retrieving its underlying data at scale and thus cannot be used for this task.

Microsoft however recently released its Academic Graph (hereafter, "MAG"), which purports to capture more than 160 million papers since 1800 and is thus similar in many ways to Google Scholar. Unlike Google Scholar, the MAG data are openly available for download by registering for an Azure account and paying the required data-transport fees (approximately $60 for a full release, according to our own billing statement). MAG is subject to the Open Data Commons (ODC-By) attribution license, which permits the creation and distribution of derivative works with acknowledgment. Thus it is possible to use MAG as the target set of scientific articles for matching against PCS, and to publish the resulting dataset.

Given that MAG is newer and less well known than Google Scholar, one may be curious as to its coverage and representativeness. A direct comparison is infeasible because Google does not permit comprehensive downloading of its data, but some scholars have verified coverage in subsets. Paszca (2016) checks for the availability of 639 randomly-selected documents, finding MAG's coverage on par with Google Scholar (76.0% vs. 76.2%) and significantly higher than Scopus (66.5%) or the Web of Science (58.8%). Hug & Braendle (2017) benchmark MAG against Scopus and the Web of Science using 91,215 verified, multidisciplinary publications from the University of Zurich's Open Archive and Repository as of October 2016. Coverage of these publications was 47.2% in WoS, 52.0% in Scopus, and 52.5% in MAG. MAG was found to be

particularly superior vs. Scopus and WoS in recalling book sections and conference proceedings, both of which are frequently cited as prior art. (Scholars in several Engineering fields publish frequently or even primarily in refereed conference proceedings.) In order to further facilitate use of MAG, we provide Digital Object Identifiers as part of our redistribution (see Appendix 3). Moreover, in case researchers are concerned that certain journals covered by MAG are less legitimate than one might find in a curated database such as Scopus or WoS, we calculated the Journal Impact Factor for every MAG paper and provide this in our redistribution (see Appendix 4). Researchers thus have the option of excluding very low impact factor journals from the set of PCS matches.

ESTABLISHING LINKAGES BETWEEN PATENTS AND SCIENTIFIC PAPERS

We link non-patent references in patents granted by the USPTO from 1947-2018 to articles captured by the Microsoft Academic Graph. We focus on citations from U.S. patents given the USPTO's requirement "to disclose to the Office all information known to that individual to be material to patentability" (see https:// https://www.uspto.gov/web/offices/pac/mpep/mpep-2000.pdf). Applicants are in a better position to know the scientific articles on which they relied than are patent examiners, who assume the burden of finding prior art in major non-U.S. jurisdictions. One would thus expect non-patent references in USPTO documents to be at once more complete and also more representative of the science upon which the inventors actually relied, as compared to jurisdictions where no such duty exists.

References can be either to patents or non-patent literature and can appear either in the body of the patent or on its "front page." We engaged two patent attorneys and two patent examiners to better understand the nature of such citations. All four described a similar asymmetry between attorneys, inventors, and examiners with regard to the types of prior art they include in patents.

Attorneys typically assemble the list of patent-related prior art but have less to add in terms of academic literature, as they are less familiar with it. (That said, attorneys may "borrow" non-patent prior art from related patents.) By contrast, inventors themselves rarely report patents that should be included in the application ("maybe one out of twenty inventors knows a relevant patent", said one attorney), but they are the primary sources of scientific references and other non-patent prior art. Importantly, the duty of the applicant to the USPTO is to report prior art that of which the applicant is *aware*; applicants are not required to do an exhaustive search.

Examiners, of course, are quite familiar with the patent corpus (and add up to 40% of patent-based prior art) but are less familiar with the scientific literature (although one attorney maintained that examiners are "getting better" at knowing relevant non-patent publications). Indeed, both examiners said that they regularly search the scientific literature using Google Scholar and similar tools in order to find relevant non-patent prior art during the examination process. That said, both examiners said that their preference is to cite patents when possible as these tend to be more precise and relevant to patentable material whereas it can be harder to pin down the exact content and its relevance to a pending application.

The attorneys' and examiners' observations are consistent both with fieldwork and the NPL corpus itself. From in-depth interviews with 21 inventors who cited scientific articles in their patents, Bikard and Marx (2018) report that most were from the inventors themselves and not from the patent attorneys, suggesting again that PCS may more authentically represent knowledge flows including from academia to industry. Ahmadpoor and Jones (2017) find that only 4% of PCS are added by patent examiners, which we confirm in our analysis.

Regarding the location of prior art, attorneys and examiners alike were sanguine with regard to the role of citations on the "front page" of the patent as opposed to in the narrative or "body"

of the document. In February of 1947, the USPTO began listing on the front page of granted patents the prior art against which the patent itself was defined as novel and non-obvious. “The patent is presumed valid over those references,” said one attorney we consulted. Meyer (2000) adds that the front-page citations may be overgenerous as applicants attempt to impress examiners with a long list of prior art against which the present invention is (supposedly) distinct. (One examiner confirmed this observation, unprompted, naming some firms that routinely include hundreds and sometimes thousands of non-patent citations.) Many of the front-page references also appear in the body of the patent, but certainly not all. Sometimes references will be only in the body of the patent “to explain well known things without having to go into gory detail…sort of a shorthand,” as one attorney said. Thus references in the body of the patent may provide additional insight into the workings of the invention and science upon which the inventors have built, independent of whether the patent’s validity depends on differentiating itself from those references. The other attorney suggested that most citations in the body of the patent ought to be incorporated in the Invention Disclosure Summary (IDS), because citations in the body of the patent will not be reviewed by examiners and thus do little to increase the patent’s chance of being granted. He suggested that body-text citations have become less common in the past ten years, especially among newer attorneys, and he advises his clients not to include citations in the body of the patent. One of the examiners similarly expressed puzzlement at the use of body-text citations: “I can’t think of any reason not to include a body-text citation in the IDS.” Both speculated that the use of body-text citations may be on the wane, though this remains an empirical question.

Almost all PCS datasets, including ours, focus on citations that can be extracted from the front page of the patent. Bryan, Ozcan, and Sampat (2019) offer a partial dataset of citations from

both the front-page and the body text of patents, linked to papers in 244 journals from 1984-2016.[1] The field still awaits a comprehensive dataset of citations from the body text of patents. From the front pages of patents, from 1947-2018 we found 36,020,060 non-patent references.

Challenges

Linking NPLs from patents to scientific articles (whether in MAG, or other dataset) includes at least three challenges:

Knowing which non-patent citations represent scientific articles. Of the ten randomly-selected non-patent references shown in Table 1, only six are to scientific articles. Two of the references are to product brochures or user manuals; one is to a patent application; and another references an action by the patent office. Other types of non-patent references include web pages, popular magazines, and lawsuit-related documents including deposition testimony. Using the count of non-patent references as an indicator of how often scientific articles are cited is thus misleading, as noted by Cassiman, Veugelers, and Zuniga (2008).

Table 1 about here

Handling incomplete references to academic articles. Even if one can determine which of the non-patent citations are to scientific articles, determining exactly which article is being cited is difficult for a number of reasons. In Table 1, journal names are frequently abbreviated (Nucleis Acids Res., JAMA, Arch Surg). The volume and issue number of the journal are not always present; often, both are missing. Or, if included, one or the other might be incorrect. Quite often, the title of the article is truncated, partially misspelled, or entirely absent. The reference may be

[1] As a benchmark, the top 250 journals in MAG from 1984-2016 contain 0.5% of all MAG articles.

to a working paper, the title of which evolves by the time the article is finally published. References are occasionally written in a different language. In some cases, even author names or year of publication can be missing or incorrect. Trying to match incomplete or incorrect citations to scientific articles can result in both Type I and Type II errors.

Computational complexity. The non-patent citations in Table 1 are sampled from 36 million non-patent references since 1947. Checking each of these against the nearly 50 million articles in the Clarivate Web of Science (WOS), or the estimated 160 million articles indexed by Google Scholar (Orduña-Malea, et al, 2014), could involve quadrillions of patent-article comparisons. The computational task is further complicated by the fact that multiple pieces of information per citation—e.g., author, year, volume, number, page, journal name, title—may need to be checked as part of each pairwise comparison.

The MAG article data are structured, with separate fields for article title, author, journal, publication year, volume, issue, and page numbers. If the non-patented references were also structured, our task would be greatly simplified as we could execute a simple database join on the same fields in both databases, possibly introducing fuzzy matching to account for typographical errors. However, as is visible in Table 1, the non-patent references are not structured consistently. Although there are some structural tendencies—e.g., author names tend to appear at the beginning of the unstructured string—such heuristics are not always reliable.

It is especially difficult to determine which (if any) part of the unstructured string contains the title. As is visible even among the ten randomly-sampled unstructured references in Table 1, the title usually but not always appears after the author. Titles are delimited by quotes in many but not all cases; sometimes, the journal name is also/instead in quotes. Titles are very often shortened and sometimes are missing entirely. Volume/issue/page information is usually present

but is often missing or only partially available and in various orderings. Given the difficulty of imputing structure to such data, we pursue a matching strategy that makes minimal assumptions regarding the structure of the reference.

Appendix 1 describes in detail the steps involved in the linking algorithm. At a high level, we first hash the unstructured source data into millions of subsets which can then be examined in parallel. Second, we execute loose, computationally-inexpensive matching to generate a large number of potential PCS linkages. Third, we apply computationally-expensive scoring techniques to determine the likelihood that each potential PCS represents an actual PCS, and assign a confidence score to the linkage.

Resulting matches

Both the PCS linkages as well as selected MAG metadata are available at **http://relianceonscience.org** with accompanying documentation. Two sets of output are available. First are PCS based on linking USPTO to MAG. Each PCS is labeled as originating from the applicant, examiner,[2] or unknown and is given a confidence score from 3-10 (matches with confidence scores 2 or 1 are not included in the distribution). The schema for this output file is detailed in Appendix 2.

Researchers interested only in the number of PCS per patent may find this sufficient; however, we suspect that most researchers will want to know about papers that were cited, as

[2] Conversations with patent examiners highlighted that even when a reference is labeled as "added by examiner" the reference may have *originally* been added by the applicant, but because the examiner chose to list it explicitly that information is lost. Whether an examiner-added citations was originally added by the applicant is captured on the 1449 table submitted by the original applicant, but these data are not publicly available in machine-readable format. The examiners speculated that this was not frequent, but the 95% of citations added by applicants may be slightly understated.

well as papers that were not cited. Hence, we post selected metadata from the 1 January 2019 edition of the Microsoft Academic Graph, including year, volume, issue, pages, title, journal / conference, authors, subjects, and citations. Importantly, these files include Digital Object Identifiers (DOIs) for all MAG papers where available and can be merged against the master file to provide a crosswalk to other databases of papers. These files and schemas are described in Appendix 3. Finally, additional fields we add to MAG including Journal Impact Factor are in Appendix 4.

Turning to the matches themselves, we provide 16,715,523 linkages from 9,472,232 unique non-patent references in 1,479,338 unique patents, citing 3,937,792 unique MAG papers. In 2006 and later, 94.9% of PCS are from applicants, and 4.9% are from examiners, consistent with the reports of others (Ahmadpoor & Jones, 2017). Note that prior to 2006, applicant/examiner indicators are rarely available.

Approximately 17.6% of patents granted since 1947 contain at least one citation to science on their front page. That trend is growing, up from 6.7% in 1976 to 25.6% in 2018.[3] Patents have on average 1.99 citations to science, a trend which has grown substantially since 1947 as depicted in Figure 1. Patents before 1980 had less than one citation to science on average, but more recently the average has been more than four citations per patent.

Figure 1 about here

Academic patents have far more citations to science than those assigned to firms (14 on average, versus 2 for corporate patents and 1.3 for government patents), a rate which has grown

[3] We found citations to science in about 1% of patents from the late 1940s and early 1950s. However, given that OCR errors make it more difficult to identify citations before 1976, those match rates may be understated.

dramatically, especially in the 1990s, as shown in Figure 2. This is consistent with the observation of one of the attorneys we interviewed, who said that the sort of academic scientists whose works are patented by universities know the academic literature extremely well and cite it generously. By contrast, inventors in firms are not as well acquainted with published work. Indeed, although there is some growth in the number of scientific citations per patent among corporations (and government), the rate of increase pales in comparison to universities. Lone inventors cite less science than any of the other groups, with less than half a citation to science per patent and little growth in this rate.

Figure 2 about here

Which fields of innovation are most reliant on science? Figure 3 shows the average number of citations to science per patent, broken down by eight primary Cooperative Patent Classification categories, which have been retrieved for patents back to 1926 by Fleming et al. (2019). Of these categories, Chemistry and Metallurgy has the highest number of citations to science per patent (average of 6 per patent, and up to 17 per patent in recent years), followed by Human Necessities. Mechanical Engineering is the least reliant technology category, with 0.14 citations to science per patent and little growth over the full timespan.

Figure 3 about here.

Approximately 1.5% of all papers are cited by the front pages of USPTO patents. By far, the journal most frequently cited by patents is Proceedings of the National Academy of Sciences, followed by the Journal of Biological Chemistry, Science, and Nature. The top 20 most cited journals are listed in

Table 2. The life sciences are very heavily represented among top journals, as is physics. The paper most frequently cited by patents (4,783 times) is “Less than additive epistatic interactions of quantitative trait loci in tomato” by Eshed, et al., published 2004 in the journal Nature. The top 20 most cited papers are in Table 3.

Table 2 about here

Table 3 about here.

Performance Characterization

When matching patents with papers at scale, some error is inevitable. The algorithm may fail to accurately map an NPL reference to the appropriate paper in MAG (false negative), or it may mistakenly map an NPL reference to the wrong paper in MAG (false positive). Which matches should be provided for use by others? Of course one would like to avoid any false positives or false negatives, but doing so would involve checking 16.7 million PCS linkages by hand and is impractical. Moreover, there is a tradeoff in that reducing false positives will increase false negatives and vice versa.

One approach is to present a single set of PCS linkages which we believe best trades off precision and recall. However, researchers may have different preferences for false negatives vs. false positives. For example, in estimating percentage of patents relying on government funding by using PCS linkages, Fleming et al., (2018) chose a conservative matching approach in the interest of constructing a lower bound. By contrast, researchers interested in using PCS linkages in a particular industry or even for a single firm, may prefer to start with a less conservative set of matches for their narrow context, perhaps checking the few applicable PCS manually.

Respecting these preferences, instead of exercising our own judgment in guessing which set of matches are most appropriate for researchers, we provide a large set of matches along with confidence scores from 3-10. Note that 87.5% of matches have confidence score = 10 (essentially error-free; see below for details of performance analysis). The percentage of matches in each confidence band is shown in Figure 4.

Figure 4 about here.

Of course, researchers need to understand what a confidence score means in terms of false positives vs. false negatives in order to make a sound decision regarding which set of matches to use. A confidence score of "9" is uninformative without further detail. In the section below, we characterize false-negatives and false-positives at each confidence level. Given that many may find it more intuitive to think in terms of correct performance instead of errors, we report performance in terms of two common metrics:

Recall: What percent of actual matches did the algorithm find? Higher is better. Given that a false negative indicates an actual match that the algorithm failed to find, recall is equivalent to 1 minus the percentage of false negatives. (A common synonym for recall is "coverage" as it represents the percentage of actual matches found by the algorithm.)

Precision: What percent of the reported matches were correct? Again, higher is better. A false positive means that a reported match was incorrect, i.e., the PCS linkage pointed to the wrong article in MAG. Thus precision is 1 minus the percentage of false negatives. (Precision is often referred to as "accuracy.")

Precision was evaluated by checking the accuracy of a stratified random sample of the paper to patent matches output by the algorithm. For each confidence level, 100 randomly-selected matches output by the algorithm are checked by hand for accuracy. A research assistant checked these independently, and then reviewed the results with one of the authors.[4] The number of false positives at each confidence level are listed in the third column of Table 4. The corresponding

[4] Note that there would be a risk of overfitting the algorithm to the matches that were reviewed manually, especially if these were used again in testing precision. Although thousands of hand-scored matches were retained and used to assess the progress of the algorithm, the precision scores in Table 4 were derived from a fresh set of 1,000 hand-checked matches.

percent correct for each confidence level is listed in column 4. This percentage is multiplied by the number of matches found at each confidence level (column 2) to estimate the cumulative percent of correct matches per confidence level (column 5).

Table 4 about here

As is visible from Table 4, at confidence levels 2 and 1 very few correct matches are likely to be found. Thus we restrict our distribution to PCS linkages at confidence score 3 and above.

To assess recall (again, 1 minus false negatives), we need to have a set of actual matches the algorithm *should* have found. We created a test set of actual, "known good" references. Of course, the algorithm developers cannot involve themselves in the creation of the known-good references, lest the algorithm be overfitted to these test cases, and the algorithm would inevitably appear to perform better (on the test data) than it does generally. Accordingly, we tasked multiple research assistants with creating the known-good cases from a random sample of 1000 unstructured NPLs. The RAs were trained by categorizing 100 randomized unstructured lines under supervision of one of the authors, but these were discarded from the test set. The authors have never seen the known-good references.[5]

The first step in creating the known-good list was to categorize the 1000 unstructured non-patent references into those that are scientific references and those that are not, as in Figure 1. Two RAs did this independently, and differences were resolved via conference, with 546

[5] One might argue that even if the algorithm developers have not seen the known-good references, even being told the performance on that test set may result in overfitting if, for example, the algorithm developers try techniques that happen to work well on this test set. That the test set was randomly sampled works against such bias, but we cannot fully rule out this possibility.

scientific references retained. Next, it was established which of these 546 scientific references were findable in MAG. The RAs jointly determined that 501 of these were in MAG.

The output of the algorithm was automatically compared against the known-good patent-to-paper references. Table 5 shows the number of known-good references found at each confidence level, individually and cumulatively. The recall % is cumulative. At the least-restrictive level of matching (1), more than 93% of known-good references were identified.

Table 5 about here

Researchers may have different preferences for recall vs. precision.

Figure 5 plots recall against false negatives (i.e. one minus precision), using the statistics from Table 4 and Table 5.[6] Recall is on the y-axis and precision on the x-axis. A "perfect" algorithm would have 100% recall (i.e., no false negatives) and 100% precision (no false positives) and thus would plot in the upper right-hand corner of the graph. The plotted points in

Figure 5 show how recall and precision can be traded off against each other. Each point on the graph represents recall and precision scores at a particular confidence score, shown italicized in parentheses.

Figure 5 about here

For instance, researchers who care about finding as many matches as possible should select confidence score 3, which is associated with 93.01% recall and 98.76% precision. That said, a substantial improvement in precision is achieved by moving to confidence score 4 (99.47%) with

[6] We thank Ivan Png for suggesting a reorientation of this graph from the typical Receiver-Operator Curve (ROC) orientation with recall on the y-axis and false positives on the x-axis. Having instead precision (1 – recall) on the x-axis makes clearer that the data user is trading off false negatives vs. false positives in deciding which confidence level(s) to use.

only a slight decrease in recall (92.81%). At the other extreme, those who want perfect or 100% precision (i.e., no false positives) could choose only matches with confidence score = 10. However, insisting on 100% precision lowers recall to 84.63%. Recall grows quickly with only a slight decrease in precision: for example, using all matches with confidence 5 and above yields 92% recall with 99.5% precision.

Comparison with prior PCS efforts is not straightforward for two reasons. First, few other PCS linkages are publicly available, and those that are have generally not provided performance metrics. Ideally we would compare our matching performance against another dataset matching NPL to MAG, but the only effort we are aware of matching patent NPLs to MAG is available via API access at lens.org (Jefferson, et al. (2018)). Jefferson et al. report that their algorithm reports matches with a confidence score of .9 or greater, but the false-negative/false-positive characteristics of this confidence level are not reported. Although our own confidence score of 9 or above corresponds to 99.96% precision and 87.62% recall, we cannot say how that compares without knowing similar metrics for other datasets. If it becomes possible to retrieve their PCS linkages with MAG IDs at scale, we would be able to automatically assess its false-negative performance against our known-good set and could assess its false-positive performance in a random sample.

As an indirect comparison, our algorithm also operates on the Web of Science (WoS) data although we do not publish the resulting matches due to licensing restrictions. Although we have not calculated formal precision and recall performance for our patent-to-WoS matches, it seems a reasonable assumption that matches with confidence = 10 will have perfect recall. Thus one method of comparing against prior efforts is to count the number of error-free PCS. Fleming et al. (2019) enable such a comparison because they report that their PCS have 100% precision and

20% false negatives in a known-good sample using WoS, with a total of 9,589,207 matches. Their sample ranges from 1976-2017, so as a comparison we count the number of patent-to-WoS matches generated by our algorithm with confidence = 10 during the same time frame. We find 10,425,123 such matches, nearly a 9% increase in the number of PCS.[7]

As a second comparison, Azoulay, Graff Zivin, and Sampat (2012) found 558,982 unique articles in PubMed that were referenced from USPTO patents in the NBER-defined Chemical and Drugs & Medical categories between 1976 and 2010. Similar to Fleming et al. (2019), they reported no false positives in a random sample of 200. We counted the number of PCS linkages from that same set of patents to WoS papers for which a PubMed ID could be found. (PubMed IDs for WoS papers were identified via a crosswalk supplied by Clarivate, publishers of the Web of Science; the accuracy of their crosswalk is unknown but presumed to be high.) We found 579,019 unique PubMed matches with confidence = 10, an increase of 3.6%.

Known Limitations

There are two types of references that will not be found via our algorithm. First, although our algorithm can find matches where the original unstructured line is missing the year, a reference containing a year that differs by more than one year (e.g., 2005 instead of 1995) will not be found. A second category is references that a) omit or misspell both the longest and second-longest

[7] Knaus and Palzenberger (2018) report extensively on their matching of WoS to USPTO, EPO, and WIPO patents, although they use a different methodology to calculate precision and recall, so their results are not directly comparable to ours. They report achieving precision of >99% for 40% of the USPTO NPLs they checked by hand, and 80% if relaxing precision to 90%. They also forecast matching approximately 9.5 million NPLs to WoS, similar to the count of Fleming et al. (2019), although it is unclear whether this includes USPTO, EPO, and WIPO patents.

words in the title, such that our loose first-pass title match will not find it, and also b) do not include the first page (or volume, if MAG is missing the first page) of the article.

Beyond these immediate issues, there are many ways to potentially enhance the performance of the algorithm. We rely on matching the first author of the paper (by surname, and where possible by first initial of the given name). Sometimes, however, the unstructured line includes not only the first initial but the entire given name, which we could use to increase our confidence in a particular match. Similarly, sometimes more than one author is listed and so we could leverage matching on multiple author names to increase confidence, especially given low title-match score. Moreover, we can take advantage of the prior probability of author X publishing a(ny) paper in year Y to adjust our confidence scoring.

Certainly the greatest limitation of this dataset is its exclusive focus on front-page citations to non-patent prior art. As Bryan et al (2019) show, there are many citations that are not on the front page but are embedded in the body text of the patent (moreover, many front-page citations also appear in the body text). Their examination of citations to 244 journals from 1984-2016 suggests that these are distinct in purpose and function from front-page citations (to the extent that the two types do not overlap). The creation of a comprehensive dataset including body-text citations from all issued patents to all known journal articles is an important next step.

CONCLUSION: HOW PCS CAN FUEL FURTHER STRATEGY RESEARCH

We have described the construction of a set of citations from USPTO patents, 1947-2018, to papers 1800-2018 from the Microsoft Academic Graph. The open nature of MAG makes it possible for us to share these patent-to-paper citations for use by other researchers. We moreover

characterize the performance of our linkage algorithm, characterizing false-positive and false-negative rates for linkages at each confidence level.

The general availability of patent citations to science will enable researchers to pursue a number of research agendas that were previously difficult to attempt, at least at scale. In the introduction we noted three areas the researchers have pursued: 1) describing the process of searching for innovations 2) characterizing the nature of R&D portfolios 3) the localization of spillovers from academia to industry. We conclude by sketching additional research questions that could be answered using PCS, including recent papers that further this agenda.

Reliable counts of scientific references in patents

Before proceeding, it is worthwhile to document what information is available to researchers in this dataset that was not generally available from the raw patent records. Previously, researchers have been able to download the non-patent references from the front page of U.S. patents in their raw, unstructured form. Absent additional processing, the general applicability of these data is to assemble a count of the non-patent references but without regard to the nature of the references (as in Fleming & Sorenson, 2004). As shown in Table 1, and as discussed more generally by Callaert et al. (2011), a large percentage of non-patent references are not scientific in nature but correspond to product brochures, actions of the patent office, press releases and other news items, or legal proceedings. Therefore, a key contribution of this dataset is to enable researchers to assemble a much more accurate count of scientific references in patents.

As one example, Kneeland, Schilling, & Aharonson (2019) use these data to examine the process by which inventors in firms come up with non-incremental innovations, finding that "outlier" patents have more citations to science. Importantly, they find stronger results when

using the count of PCS from these data than when simply counting the overall number of NPL citations. Arora, Belenzon, and Suh (2019) use these data to expand the concept of firms' technological search into markets for technology. Beyond merely citing articles they find relevant, firms can engage in transactions to acquire intellectual property (i.e., patents) they deem valuable. They report that patents more reliant on science (as measured by the count of PCS) are considerably more likely to be traded than those lacking scientific references. Two separate teams of authors use these and similar PCS to associate the count of scientific references in a focal patent with the monetary value of a patented invention (Watzinger & Schnitzer, 2019; Poege et al., 2019). Whereas these authors would have had to use simple counts of NPL citations, their results are more reliable because they use actual counts of PCS.

Counts of scientific references from patents to a focal article

Not only does this dataset enable a more true count of citations to science in a patent; it enables the researcher to know how many citations to a focal paper come from the front pages of patents. Such analysis is impossible with a raw count of NPL citations, even if one decomposes NPLs into scientific and non-scientific. Rather, one must link the scientific NPL citation to the actual paper. The availability of these linkages enables research previously only possible for research teams that invested time in constructing the patent-to-paper linkages.

As an example, Bikard & Marx (2019) deepen our understanding of biases and heuristics involved in firms' search for external technologies by counting the number of patent citations to each of more than 10 million articles. They find that firms tend to pay more attention to papers located in "hubs" of relevant industry R&D. They moreover characterize papers as being in more vs. less applied fields by counting the average number of PCS to all papers in the same field as the focal paper. Finally, the introduce a measure Journal Commercial Impact Factor (JCIF),

similar to the traditional Journal Impact Factor but which characterizes the commercial influence of a paper by counting citations from patents.

Enabling explicit comparisons between linked patents and papers

Beyond creating counts of scientific references in a given patent, or counts of scientific references to a given paper, these data enable explicit comparisons between each linked paper and patent. Such comparisons can be drawn across a wide variety of characteristics, including geography, age, institution, status, and other factors.

For example, Watzinger et al. (2019) take advantage of the ability to draw geographic comparisons between citing patents and cited articles to predict the impact of hiring a new professor on local commercial activity. They count patent citations to the papers published by a newly-hired professor—*after* s/he was hired—as compared to other shortlisted candidates who were not hired. By measuring the proximity of each citing patent to the cited paper, they are able to count the number of PCS that occurred in the local area of the focal university and distinguish this count from more distant citations. Beyond microgeography, opportunities abound to understand cross-state or cross-country patterns of reliance on science.

Marx and Hsu (2019) leverage authorship comparison facilitated by this dataset to catalogue Science-Based Ventures (SBVs) in which a startup company founded by a university scientist commercializes that same discovery. Each patent assigned to a startup (as determined from CrunchBase or VentureXpert) is compared with all of the papers it cites to determine the level of overlap between the authors of the paper and the inventors on the patent. The count of SBVs in North America closely parallels the counts reported by the Association of University Technology Managers, but their data are available worldwide.

Many additional such comparisons are possible. Another application of authorship comparison would be to correlate the relative status of the paper cited (i.e., by the authors' collective citation history vs. that of the patent inventors). One could alternatively rank the scientific status of a patent according to the average JIF of cited articles, average number of citations to cited articles, or other measures. Do firms rely on the "usual suspects" when citing scientific literature, or do they unearth less-well-trodden discoveries, and how does this practice relate to the novelty of the patented invention?

In addition to comparing authorship, one can compare institutional affiliation of the paper with that of the patent assignee (noting, of course, that papers may have multiple affiliations and patents may have multiple assignees). Doing so may be a useful bookkeeping exercise for researchers who want to exclude "self-citations"; more broadly, however, differentiating internal vs. external citations can open a window into how insular firms' reliance on science is. Young firms likely have little internal science to rely upon, but does this pattern change as firms age? What explains which firms continue to absorb external knowledge with those that focus inward, and do these practices predict differences in market performance?

Likewise, what leads firms to rely on older vs. more recent science in their R&D process? Comparing the time lag between the citing patent and the cited paper could afford insight into firms' and inventors' preferences for well-established vs. cutting-edge science. Individual authors or inventors' mobility, including exogenous limits on their mobility such as non-compete agreements (Marx, Strumsky, & Fleming, 2009), could be used to assess the impact on science reliance by the influx of "new blood" into the R&D staff of a firm.

Finally, comparisons can be drawn by field. Which inventors cite science from their own field, and who recombines scientific inventions more broadly? Does the nature of creativity

impact firm performance or the careers of inventors? To this end, some sort of crosswalk between patent classifications and paper categories must be assembled.

Closing remarks

Work to date may only scratch the surface of possible paths utilizing PCS. Invention occurs in two largely distinct, yet somewhat overlapping spheres: the practice of "open science" predominantly in academia; and the commercial realm, in which temporary monopolies can be secured. These worlds increasingly overlap as firms not just patent but publish, and as universities lay legal claim to the allegedly-open creations of their employees. (One might observe, somewhat ironically, that much if not most "open" science is paywalled and the review process fully opaque whereas the patenting process—at least in the U.S.—is exceedingly transparent with all documentation freely available.) Exploring the relationship between these spheres of invention can be challenging because publishing and patenting involve separate work products and, largely, separate actors. Perhaps the most important promise of PCS is to bridge those spheres by linking patents to papers and inventors to authors. What is the topology of the cross-community networks of authors who patent and inventors who publish? Which ideas originate in academia and then migrate to industry, and when does the reverse process take place? Our hope is that a broad set of researchers can attack these and other questions now that they are more easily able to assess reliance on science.

REFERENCES

Agarwal, S. M. Lincoln, H. Cai, and V. Torvik. Patci – a tool for identifying scientific articles cite by patents. Poster presented at the Illinois Graduate school of Library and Information Science.

Ahmadpoor, Mohammad, and Benjamin F. Jones. 2017. "The Dual Frontier: Patented Inventions and Prior Scientific Advance." Science 357 (6351): 583–87.

Arora, A., Belenzon, S., and Suh, J. (2019). Science and the Market for Technology.

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate R&D. Strategic Management Journal, 39(1), 3-32.

Arora, A., Belenzon, S., & Sheer, L. (2017). Back to Basics: Why Do Firms Invest in Research? (No. w23187). National Bureau of Economic Research.
Arts, S., & Fleming, L. (2018). Paradise of novelty—or loss of human capital? Exploring new fields and inventive output. Organization Science, 29(6), 1074-1092.
Azoulay, P., Zivin, J. S. G., & Sampat, B. N. (2011). The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine. NBER Chapters, 107–155.
Belenzon, Sharon, and Mark Schankerman. "Spreading the word: Geography, policy, and knowledge spillovers." Review of Economics and Statistics 95.3 (2013): 884-903.
Bikard, Michael and Matt Marx. (2019) "Hubs as lampposts: Academic location and firms' attention to science." Forthcoming, Management Science.
Branstetter, L., & Kwon, H. (2004). The restructuring of Japanese research and development: The increasing impact of science on Japanese R&D. Unpublished manuscript.
Cassiman, Bruno, Reinhilde Veugelers, and Pluvia Zuniga. 2008. "In Search of Performance Effects of (in)Direct Industry Science Links." Industrial and Corporate Change 17 (4): 611–46.
Callaert, Julie, Maikel Pellens, and Bart Van Looy. 2014. "Sources of Inspiration? Making Sense of Scientific References in Patents." Scientometrics 98 (3): 1617–29.
Fleming, L., H. Greene, G. Li, M. Marx, and D. Yao, 2018. "U.S. Innovation Relies Increasingly on Government Funding."
Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. Strategic Management Journal, 25(8-9), 909-928.
Gaetani, Ruben, and M. Li Bergolis. "The economic effects of scientific shocks." Unpublished Manuscript (2015).
Gittelman, M., & Kogut, B. (2003). Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. Management Science, 49(4), 366-382.
Hu, D., Chen, H., Huang, Z., & Roco, M. C. (2007). Longitudinal study on patent citations to academic research articles in nanotechnology (1976–2004). Journal of Nanoparticle Research, 9(4), 529-542.
Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. Scientometrics, 113(3), 1551-1571.
Jefferson, O. A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., ... & Bilder, G. (2018). Mapping the global influence of published research on industry and innovation. Nature biotechnology, 36(1), 31.
Katila, R. and Ahuja, G., 2002. Something old, something new: A longitudinal study of search behavior and new product introduction. Academy of management journal, 45(6), pp.1183-1194.
Kneeland, M, M. Schilling, and B. Aharonson (2019). Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation. Forthcoming, Organization Science.
Knaus, J., & Palzenberger, M. (2018). PARMA. A full text search based method for matching non-patent literature citations with scientific reference databases. A pilot study.
Lemley, Mark A., and Bhaven Sampat. 2011. "Examiner Characteristics and Patent Office Outcomes." Review of Economics and Statistics 94 (3): 817–27.
Li, Danielle, Pierre Azoulay, and Bhaven N. Sampat. 2017. "The Applied Value of Public Investments in Biomedical Research." Science 356 (6333): 78–81.
Marx, M. and D. Hsu. 2019. "The Entrepreneurial Commercialization of Science: Evidence from "Twin" Discoveries."
M. Marx, D. Strumsky, and L. Fleming, "Mobility, Skills, and the Michigan Non-compete Experiment." Management Science 55(6):875-889 (lead article) (2009).
Narin, F. and D. Olivastro (1998). Linkage between patents and papers: an interim EPO/US comparison. Scientometrics 41:1-2:51-59.
Orduña-Malea, Enrique, Juan Manuel Ayllón, Alberto Martín-Martín, Emilio Delgado López-Cózar. "About the size of Google Scholar: Playing the numbers." Mimeo, 2014.

Patelli, A., Cimini, G., Pugliese, E., & Gabrielli, A. (2017). The scientific influence of nations on global scientific and technological development. Journal of Informetrics, 11(4), 1229-1237.
Paszcza, B. (2016). Comparison of Microsoft academic (graph) with web of science, scopus and google scholar (Doctoral dissertation, University of Southampton).
Poege, F., D. Harhoff, F. Gaessler, and S. Baruffaldi (2019). "Science Quality and the Value of Inventions."
Ribeiro, L. C., Kruss, G., Britto, G., Bernardes, A. T., & e Albuquerque, E. D. M. (2014). A methodology for unveiling global innovation networks: patent citations as clues to cross border knowledge flows. Scientometrics, 101(1), 61-83.
Roach, Michael, and Wesley M. Cohen. 2013. "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research." Management Science 59 (2): 504–25.
Scherer, F. M. (2001). The innovation lottery. Expanding the Boundaries of Intellectual Property: Innovation Policy for the Knowledge Society, 3(3).
Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Research Policy, 46(8), 1416-1436.
Watzinger, Martin, and Monika Schnitzer (2019). "Standing on the Shoulders of Science". Mimeo.
Watzinger, Martin, Lukas Treber, and Monika Schnitzer (2019). "Universities and Science-Based Innovation in the Private Sector". Mimeo.

Table 1: Ten randomly-sampled non-patent references from the front page of U.S. patents.

Patent #	Unstructured reference	PCS
6223284	Compaq Computer Corporation, "Compaq product information, bulletin, Proliant family of servers section 8,.Copyrgt. 1994 Compaq Computer Corporation, Feb. 2, 1995, pp. 1-6.	N
9834791	Eisenschmidt et al., "Developing a programmed restriction endonuclease for highly specific DNA cleavage," Nubcleic Acids Res., 33(22):7039-47 (2005). cited by applicant.	Y
8009111	John P. Gianvittorio and Yahya Rahmat-samii, "Fractal element antennas: a compilation of configurations with novel characteristics," IEEE, 4 pages, 2000. cited by other.	Y
9113925	Dald et al., "Accidental burns", JAMA, Aug. 16, 1971, vol. 217, no. 7, pp. 916-921. cited by applicant.	Y
9782195	"Fenestration revisited"", John A. Elefteriades, MD, et al.--Arch Surg--vol. 125--Jun. 1990--pp. 786-790. cited by applicant.	Y
5383140	"User's manual, four-bit microcontroller and peripheral memory, tlcs-47e/47/470/470a"" (portions of title are in the Japanese language), pp. 5-211 through 5-223 and unnumbered final page, published by Toshiba corporation, dated 1991.	N
D699952	US. appl. no. 13/783,109, filed Mar. 1, 2013, Yang et al. cited by applicant.	N
9484093	response to office action dated Aug. 5, 2016 in U.S. appl. no. 14/715,586. Cited by applicant.	N
9518078	Wolff, Manfred e. ""Burger's medicinal chemistry, 5ed, part i"", John Wiley & Sons, 1995, pp. 975-977. cited by examiner.	Y
8980864	Saenz-Badillos, J. et al., RNA as a tumor vaccine: a review of the literature. Exp Dermatol. jun. 2001;10(3):143-54. cited by applicant.	Y

Table 2: The top 20 journals by the number of PCS.

# PCS	*Journal*
428,363	Proceedings of the National Academy of Sciences of the United States of America
289,561	Journal of Biological Chemistry
284,282	Science
260,491	Nature
168,345	Journal of the American Chemical Society
166,660	Applied Physics Letters
130,298	Nucleic Acids Research
123,565	Journal of Medicinal Chemistry
120,647	Journal of Immunology
115,863	Cancer Research
108,815	Cell
90,148	Biochemistry
86,194	Journal of Organic Chemistry
84,344	Nature Biotechnology
83,264	Analytical Chemistry
81,706	Journal of Virology
80,432	Blood
67,249	Biochemical and Biophysical Research Communications
64,387	Journal of Applied Physics

Table 3: The top 20 papers most frequently cited by USPTO patents.

# PCS	Title	First author	Journal	Year
4,783	less than additive epistatic interactions of quantitative trait loci in tomato	y eshed	Genetics	1996
4,601	linkage disequilibrium and fingerprinting in sugar beet	t kraft	Theoretical and Applied Genetics	2000
4,242	substitutions	james u bowie	Science	1990
3,754	single amino acid substitution altering antigen binding specificity	stuart rudikoff	Proceedings of the National Academy of Sciences of the United States of America	1982
3,388	specificity	g kohler	Nature	1975
3,070	room temperature fabrication of transparent flexible thin film transistors using amorphous oxide semiconductors	kenji nomura	Nature	2004
2,879	semiconductor	kenji nomura	Science	2003
2,829	their electrical properties	satoshi masuda	Journal of Applied Physics	2003
2,829	polymer stabilized liquid crystal blue phases	hirotsugu kikuchi	Nature Materials	2002
2,827	temperature	e fortunato	Applied Physics Letters	2004
2,822	m 3 4 and 5 ingao3 zno 3 and ga2o3 zno m m 7 8 9 and 16 in the in2o3 znga2o4 zno system	noboru kimizuka	Journal of Solid State Chemistry	1995
2,812	carrier transport in transparent oxide semiconductor with intrinsic structural randomness probed using single crystalline ingao3 zno 5 films	kenji nomura	Applied Physics Letters	2004
2,795	field effect transistor on srtio3 with sputtered al2o3 gate insulator	kazuo ueno	Applied Physics Letters	2003
2,794	modulated structures of homologous compounds inmo3 zno m m in ga m integer described by four dimensional superspace group	chunfei li	Journal of Solid State Chemistry	1998
2,791	irradiation with ultraviolet lamp	naoko asakuma	Journal of Sol-Gel Science and Technology	2003
2,789	transistors	kenji nomura	Japanese Journal of Applied Physics	2006
2,786	42 1 invited paper improved amorphous in ga zn o tfts	ryo hayashi	Society for Information Display International Symposium	2008
2,783	dry etching of zno films and plasma induced damage to optical properties	jun-beom park	Journal of Vacuum Science & Technology B	2003
2,781	a ferroelectric transparent thin film transistor	mwj menno prins	Applied Physics Letters	1996

Table 4: Precision (1 – false positives) in a random sample of 100 PCS linkages per confidence level.

(1)	(2)	(3)	(4)	(5)
actual matches		sample of 100		precision
confidence level	non-patent references linked	manually marked incorrect	% correct	estimated cumulative % correct
10	14,632,844	0	100%	100.00%
9	653,258	1	99%	99.96%
8	404,045	3	97%	99.88%
7	292,615	7	93%	99.76%
6	155,446	11	89%	99.65%
5	172,955	12	88%	99.53%
4	112,732	9	91%	99.47%
3	291,628	41	59%	98.76%
2	379,671	79	21%	97.04%
1	589,531	96	4%	93.93%

Table 5: Recall (1 – false negatives) as measured against 501 known-good references

confidence level	non-patent references linked	# found (of 501 known)	recall
10	14,632,844	424	84.63%
9	653,258	439	87.62%
8	404,045	445	88.82%
7	292,615	455	90.82%
6	155,446	456	91.02%
5	172,955	461	92.02%
4	112,732	465	92.81%
3	291,628	466	93.01%
2	379,671	467	93.21%
1	589,531	468	93.41%

Figure 1: Average number of citations to science per patent, by grant year

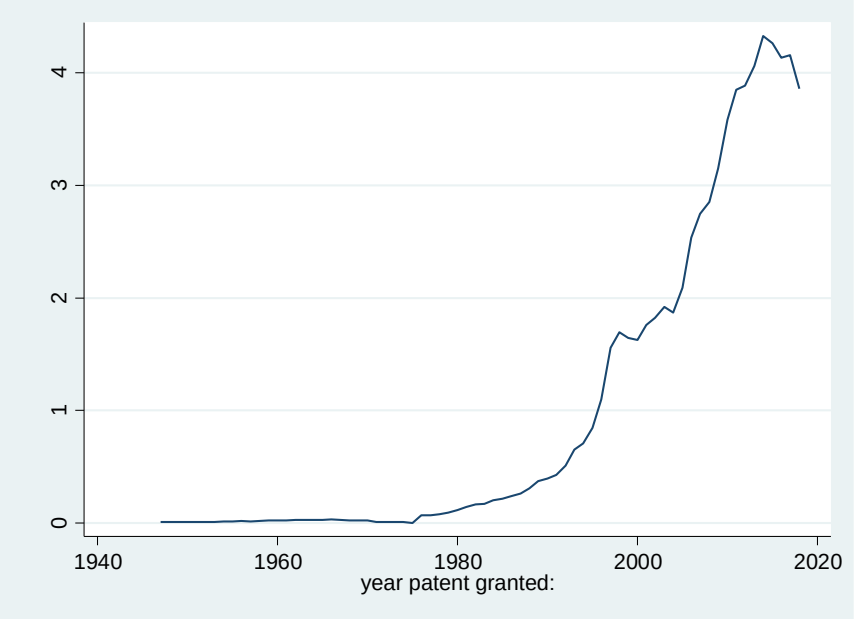


Figure 2: Average number of citations to science per patent, 1947-2018, broken down by assignee type.

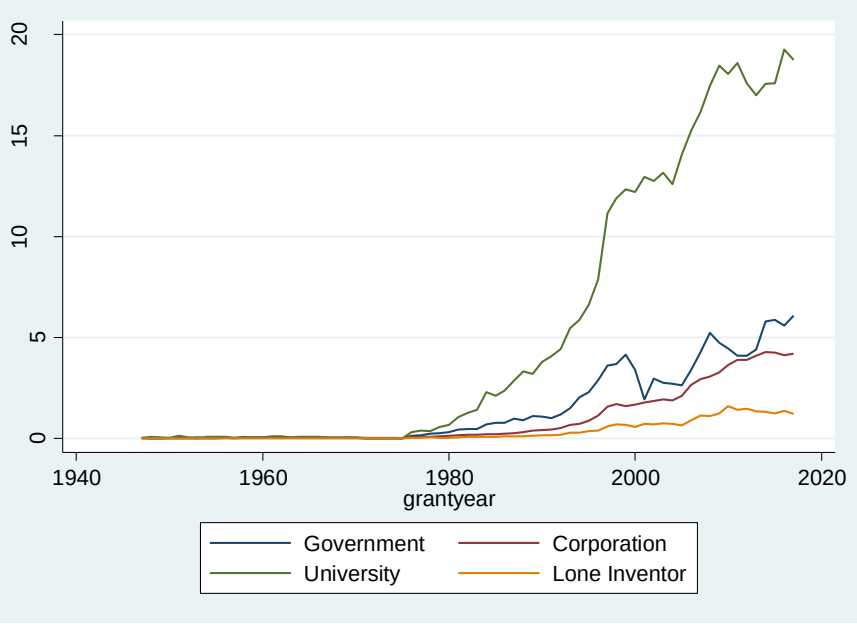


Figure 3: Citations to science per patent, by patent grant year and technical classification

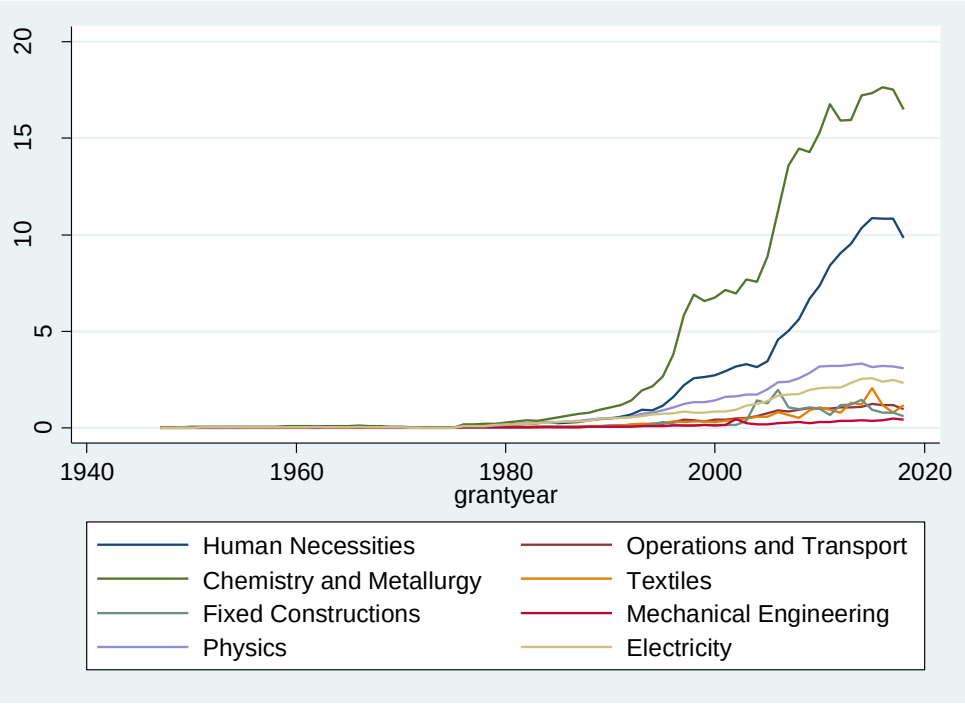


Figure 4: Distribution of confidence scores for PCS linkages

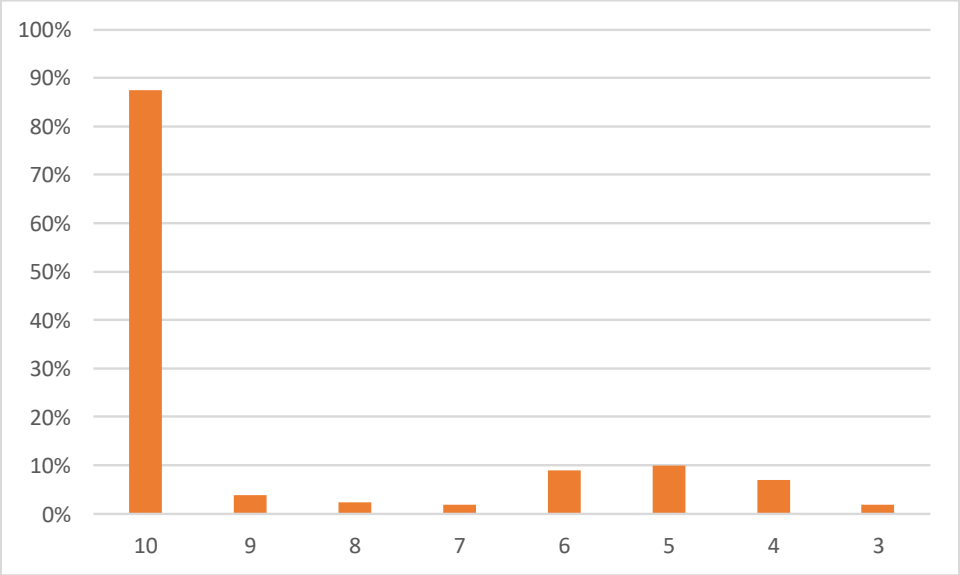
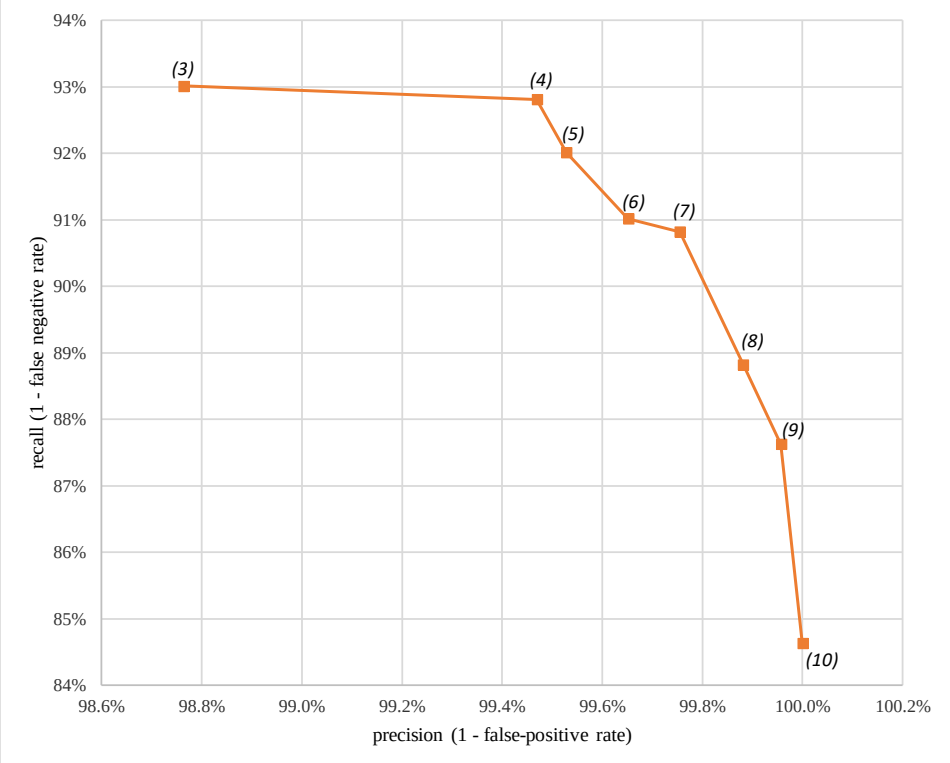


Figure 5: PCS linkage algorithm performance, recall vs. precision



Appendices for Reliance on Science in Patenting:

USPTO Front-Page Patent Citations Since 1947

Matt Marx & Aaron Fuegi

Appendix 1: Details of algorithm implementation

First, we standardize lexically both the structured and unstructured source data for analysis. Second, we hash the unstructured source data into millions of subsets which can then be examined in parallel. Third, we execute loose, computationally-inexpensive matching to generate a large number of potential PCS linkages. Fourth, we apply computationally-expensive scoring techniques to determine the likelihood that each potential PCS represents an actual PCS, and assign a confidence score to the linkage. The following sections describe each step in turn.

Step 1: Lexical standardization of structured and unstructured input data

The unstructured non-patent references requires preprocessing primarily for pre-1976 records, due to errors in optical character recognition (OCR). There are approximately 100,000 non-patent references captured by OCR, too many to correct manually without a substantial investment. Instead, we adjusted the source data in ways that could potentially be handled by the more computationally-expensive matching described below. Occasional letter substitutions will already be handled, but since our algorithm separates words based on nonalphanumeric characterics, we addressed two common errors. First, letters (especially 'a') were frequently substituted by '@', which caused words to be split and thus not match. We therefore replace '@' with 'a' when it is embedded within a word, such as "self-driving c@r" or "electr@chemical reaction" (note: flexible matching will allow for 'electrachemical' to be properly matched against 'electrochemical' even though it has been incorrectly replaced here).

Second, words are frequently split by OCR in two with a spurious hyphen, sometimes followed by a space. For example, "parametric" might be rendered as "param-etric" or "param- etric". Approximately one-third of the 100,000 pre-1976 non-patent references have words in this format. Of course, many words (and scientific words in particular) are separated by hyphens, such as "self-driving" and thus we do not want to introduce errors by falsely dropping hyphens to create "selfdriving." Thus we only recombine words separated by a hyphen (with optional trailing space) when neither of the hyphen-separated words is in the dictionary.

The structured data from MAG require less lexical preprocessing. Each is run through an ASCII filter to match the character set of the unstructured non-patent references, including the transliteration of Greek characters common in scientific titles. Articles missing authors (or where the author is listed as "Anonymous") are dropped.

Step 2. Hashing unstructured USPTO source data

As noted above, direct comparison of approximately 36 million non-patent references[8] with 150 million MAG articles would require quadrillions of pairwise comparisons. Following Gaetani and Bergolis (2015) as well as Knaus & Palzenberger (2018), we partition the matching task initially by comparing only MAG papers from a given year with unstructured references that include that same year. Thus we segment the database of unstructured references into one section for each potential article year, 1800-2018. Again, recalling that the unstructured references do not have a defined year field, it is possible that four consecutive digits in an unstructured reference are the year of the article, a page number, or a part of the title. If an unstructured reference contains more than one string of digits from 1800-2018, a copy of that unstructured reference is placed in multiple segments.

Segmenting the matching space by year reduces the number of required comparisons by several orders of magnitude, and we achieve even more dramatic improvement by further hashing the search process on other components of the unstructured lines. The annual data subsets for each of 1800-2018 are further hashed, generating a subset for each non-stopword alphanumeric string in the unstructured lines for that year. As an example, for the following reference:

> Weisenschmidt et al., "Developing a programmed restriction endonuclease for highly specific DNA cleavage," Nucleic Acids Res., 33(22):7039-47 (2005). cited by applicant.

We make 19 copies of this file, one for each word that is not a stopword like 'for' or 'and'. Thus we add a copy of this reference to Weisenschmidt.txt, Developing.txt, Programmed.txt…33.txt, 22.txt, 7039.txt…and so on. We then add copies of every reference containing the word "Weisenschmidt" to Weisenschmidt.txt, and so on. This enables our matching algorithm to look for papers by Weisenschmidt just in Wisenschmidt.txt instead of searching the entire corpus of references.

This approach may seem wasteful, given that each unstructured reference l is duplicated N times where the number of non-stopword alphanumeric strings is given by N_l. However, disk space is inexpensive compared to computational savings achieved by searching only specific sub-databases for matches as opposed to searching the entire database, or annual slices.

Step 3: "Loose" matching to generate candidate PCS

With our file-based hash table in place, we can execute massively parallelized, targeted searches for specific strings within subsets of the master database of unstructured lines. Still, some of the files are very large. Rather than attempt expensive matching on all available criteria (title, author, journal, volume, page, issue), we apply a loose matching filter as a first stage in order to generate a set of *potential* matches which can then be examined in more detail.

---

[8] As a final preprocessing step, we excluded non-patent references clearly not to scientific articles. These include office actions or patent searches, deposition testimony, etc. Screening these reduced the set of non-patent references from 36,020,060 to 26,028,093.

What sort of "loose" matching is useful to generate a set of candidate matches? Most of the unstructured strings contain the author and year of the publication, so we could consider matching simply on those two fields, but this would result in many billions of potential PCS matches. To these we add one additional field to winnow down the set of candidate matches without overcomplicating the search string, in two varieties. First, we perform loose matching adding the *longest word in the title* from MAG, in addition to the year and author surname. We also match on the *second longest word in the title* to handle cases of typographical differences in the longest word. Second, we repeat the loose matching, instead adding the starting-page number (or, if missing in MAG, the volume number) to the author surname and year. Note that these are *unstructured* searches: the year, author surname, and either longest title words or page/volume can appear anywhere in the unstructured reference.

Sometimes the first author's name is incorrectly specified in the unstructured patent reference, which jeopardizes our loose-matching scheme. In addition to an exact match on author name, we perform a flexible match using Levenshtein distance = 1 as our constraint. Given that flexible matching is very expensive at this early stage, we limit flexible name matching to the first four words of the unstructured text string, only checking these words if they are a) at least four letters long b) no more than one letter longer or shorter than the author's surname c) not preceded by "et al." which generally indicates the end of the author list.

Sometimes, the year is misspecified or missing. When misspecified, it is usually the previous or subsequent year (i.e., the reference says 1995 when the paper was published in 1994). Hence, we allow for the year to be off by one in our first-stage "loose" search. Such flexibility is also useful when the patent applicant cites a working paper which is then published in the following year. In about 5% of non-patent references, the year of the article to which it refers is missing entirely and cannot be handled as above. We collect unstructured non-patent references that lack any four-digit string corresponding to a year from 1800-2018 and match these on author name and either longest word (or second longest word) or page number (or volume if page number is missing). (Obviously, this approach results in substantial overgeneration of possible matches.)

Finally, we construct a list of potential matches for which neither any year nor any author matches but where a string of words is contained in quotes (possibly indicating a title). We then extract the string of words contained in quotes and perform a fuzzy match against all MAG articles. These are then added to the list of potential matches.

The various loose searches yield more than 2 billion *potential* PCS linkages. This is far in excess of the 36 million unstructured references in the source data and largely due to overmatching of year, surname, and page/volume. For example, MAG has more than 11,000 articles in 2015 by "Smith," so many of these will match even with a page-number restriction.

Step 4: Scoring of "loose" matches

Having generated a set of potential PCS, our final task is to apply more sophisticated (and computationally intensive) techniques to exclude false positives, based on a number of heuristics. The general shape of the scoring algorithm is detailed below, and the exact thresholds and terms are available in the posted code.

Scoring first-author name

Most candidate matches have overlapping years and author names, but some author names are more common than others. We downweight our confidence in the match for authors whose surnames a) are, composited together, the authors of more than one tenth of a percent of all articles, b) resemble month abbreviations (e.g., Jan, Jun), c) are frequent *given* names (e.g., Anthony, Morgan), d) are common terms in scientific articles (e.g., Power, Diamond), or e) consist of only two letters. Fuzzy matches, where the author surname was not an exact match, from our first round are also penalized slightly.

We also penalize potential matches where the first initial of the author does not match that found in the unstructured line. Of course, it is not straightforward to determine the first initial in the unstructured text, so we apply this test only in the cases where the author's surname appears among the first five words in the unstructured line. If so, we rely on cues including "et al" as well as "and" (either following or preceding the surname) to determine the first initial. In many cases, such as "Smith, et al." no first name is available and so this filter cannot be applied.

Scoring article title

Title scoring proceeds as follows. We break apart the structured article title into its component alphanumeric strings (words). We then look for these words in the unstructured reference and, when found, note the position or "offset" of each vs. the matching word in the structured article title. The most frequent offset among all words in the title is designated as the most likely start of the title in the unstructured string. We then again compare each word in the structured title to what we believe is the matching word in the unstructured data, based on the offset value. We also look at the words just before and after in case an extra word was mistakenly added or removed in the unstructured title.

The overall score for title similarity is determined based primarily on 1) the full number of words in the structured title 2) the number of those words that matched exactly to their corresponding word in the unstructured data 3) the number of those words that matched with only a single-letter change (i.e., Levenshtein distance 1). Matching of common words is discounted. In effect, the title score increases for a higher percentage of words in the title, and the longer the title is. Titles of fewer than five words are given less weight while titles of seven or more words that match closely have greater influence on the confidence score.

Often, what appears to be the article title is enclosed in quotes. Note that this is far from always the case; many NPL entries do not contain any quotes, and some surround journal names or other extraneous text in quotes. If, however, we find any text contained in quotes, we compute the Levenshtein distance between the text in quotes and the actual title in MAG. (If there are multiple groups of quote-surrounded text, we try all of these in turn.) Note that this approach is far from foolproof, as titles within quotes are often abbreviated (e.g., "properties of gallium arsenide…an early test"). The title score generated above as well as this title score when quotes are available are both used to score potential matches, with an item being considered a likely match if it scores highly using either method.

Scoring volume, issue, and pages

We then score the match for information other than the title. We generally refer to these characteristics as "VIP" for Volume/Issue/Pages. Our original approach with non-title matches followed Fleming et al. (2018) in requiring the 3-tuple of volume, issue, and first page in order. Such an approach generates few if any false positives but results in a large number of false negatives because many unstructured non-patent references omit the issue number, and some have only the page numbers. We give credit for matching volume, issue, or page anywhere in the unstructured string; however, titles sometimes contain numbers which could yield stray matches, especially single-digit numbers. Hence we increase confidence only slightly when single-digit numbers (esp. 1) match; matches of multi-digit numbers bolster confidence.

Confidence increases dramatically if VIP information is found in sequence, such as <volume>-<page> and especially <volume>-<issue>-<first page>-<last page>, especially when all of the VIP components are three or more digits. Confidence is boosted if these sequences are preceded by *Vol.* or when *p.* or *pp.* precedes the page number. Having both first and last page number in a sequence is especially advantageous, including when the final page number is often abbreviated to contain only digits that distinguish it from the initial page number (e.g., "255-73").

By the same token, if *Vol., p*., etc is followed by a number that does *not* match the structured data, we penalize the confidence score. Moreover, if in the unstructured string we see what appear to be a volume-issue-page combination, or two page numbers in a row, but these do not match the data in MAG, we lower the confidence score. Note that this filter is not applied if both numbers in the <first page>-<last page> sequence are lower than 32, which may indicate a date range for a conference.

Scoring journal names

We increase our confidence score if the journal title is found in the unstructured string. Journal titles are frequently abbreviated in references, so in addition to searching for the canonical journal name listed in MAG we also search for shorthand versions of every journal name based on the concordance found at https://images.webofknowledge.com/images/help/WOS/A_abrvjt.html. In addition, we reviewed thousands of randomly-sampled outputs labeled correct but which did not have a match on journal to find additional abbreviations. (Proceedings of the National Academy of Sciences USA had more than three dozen abbreviations.) We give less credit for finding journal names that are common words in articles, such as "Science" or "Cell."

A composite confidence score is then determined based on the above scoring algorithm. These scores vary according to the fuzzy-match title score, journal score, and the completeness of the volume/issue/page match. Note that there may be more than one MAG ID found for a given patent/NPL combination. In such a case, we pick the MAG ID with the highest overall confidence score (or, if multiple matches have a similar overall confidence score, we pick the match with the highest title score (and further break ties with VIP score).

Appendix 2: Schema for patent citations to science (PCS) output files

The main output file, available at **http://relianceonscience.org**, is called *pcs.tsv* and is a tab-separated file containing the patent number, the unique identifier in the MAG database, confidence score, and whether the reference was filed by the applicant, an examiner, or other (if known). It contains PCS links of confidence score 3 or higher. Those using this data are asked to cite this paper. The schema is as follows:

Table A1.1: Contents of *pcs.tsv*.

Variable	Type	Notes
reftype	string	App = from applicant Exm=from examiner Unk = if unspecified in the unstructured reference (Note: almost every reference before 2006 is of unknown origin.)
confscore	numeric	Assigned confidence score to the match. Note that only matches with a confidence score of 3 or above are included in the distribution.
paperid	numeric	Unique identifier for each paper in the Microsoft Academic Graph.
patent	string	Patents are 1947-2018, granted by USPTO. Not all patents contain references to science. Only patents for which our algorith established a PCS linkage are included.
nplwithoutpatent	string	Unstructured reference to non-patent literature (NPL) from the patent. May have slight formatting alterations from original USPTO, but alphanumerics should be identical. Lowercase.

As described in the body of the paper, PCS are established via a probabilistic algorithm. Users of the data should consult Tables 2 and 3 as well as Figure 1 to determine their desired confidence-score cutoff. Matches for confidence scores 2 and 1 are not included in the distribution as there are very few correct matches at those levels. Even at confidence score 3, about half of the matches are incorrect. Most users will want to only use matches with a score of 4 or higher.

Appendix 3: Files for Microsoft Academic Graph metadata

Also available is a series of files with metadata regarding not just the references reported in Appendix 1 but *all* papers in the 1 January 2019 release of the Microsoft Academic Graph (MAG). They are compressed using the 'zip' utility under Unix CentOS5. Reposting of these data is facilitated by the ODC-By license (https://opendatacommons.org/licenses/by/1-0/index.html), under which MAG is provided and under which these data are also provided. Those using these data should cite the following paper: *Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246.*

Researchers who prefer to download the original MAG data directly from Microsoft can do so by signing up for an Azure account and billing plan, contacting Microsoft for access to MAG, selecting the 2019-1-1 release, and downloading the desired files. Instructions are at https://docs.microsoft.com/en-us/academic-services/graph/. Note however that some of the original MAG files are several dozen gigabytes in size; for example, the Papers.txt file from which several of these files are derived, is 56 gigabytes.

All files are in tab-separated format, compressed as .zip files. The first set of files contain direct metadata for papers in MAG.

Table A2.1: Contents of files with direct MAG metadata

Filename	Variables	MAG file (fields)	Notes
paperyear	paperid, paperyear	Papers.txt (1,8)	
papervolisspages	paperid, papervolume, paperissue, paper1stpage, paperlastpage	Papers.txt (1,14,15,16,17)	Issue and pages are sometimes blank. First page is available more often than last page.
papertitle	paperid, papertitle	Papers.txt (1,5)	Titles are often blank for conference papers.
papercitations	citingpaperid, citedpaperid	PaperReferences.txt (1,2)	Adds headings to PaperReferences.txt.
paperdoi	paperid, doi	Papers.txt (1,3)	DOI is not available for every paper in MAG
paperauthororder	paperid, authorid, authororder	PaperAuthorAffiliations.txt (1,2,4)	Author order not available for every author
paperauthoraffiliationname	paperid, authorid, affiliationname	PaperAuthorAffiliations.txt (1,2,5)	Affiliation not available for many authors

The next set of files contain indirect metadata, i.e. identifiers that need to be matched to dictionaries in the next set of files. One could provide the full strings of the authors, journals, etc., directly but the files would be much larger and unnecessarily redundant.

Table A2.2: Contents of files with indirect MAG metadata

Filename	Variables	MAG file (fields)	Notes
paperconferenceid	paperid, conferenceid	Papers.txt (1,13)	
paperfieldid	paperid, fieldid	PaperFieldsOfStudy.txt (1,2)	ID for field of paper.
paperjournalid	paperid, journalid	Papers.txt (1,11)	

The third set of files contains the string values for indirect metadata identifiers:

Table A2.3: Contents of files with string values for indirect MAG metadata

Filename	Variables	MAG source (fields)	Notes
authoridname_normalized	authorid, authorname_normalized	Authors.txt (1,3)	Lowercase name w/o punctuation.
authoridname_raw	authorid, authorname_raw	Authors.txt (1,4)	As originally appeared.
conferenceidname	conferenceid conferencename	ConferenceInstances.txt (1,2)	Name of conference
fieldidname	fieldid fieldname	FieldsOfStudy.txt (1,3)	Paper field, inferred from title+abstract.
journalidname	journalid journalname journalissn	Journals.txt (1,3,5)	ISSN is often unavailable.

Appendix 4: Schema for extensions to the Microsoft Academic Graph (MAG) data

In addition to the redistribution of the MAG data, we provide two extensions for fields not present in the MAG data. First, we calculate Journal Impact Factor for all journals in MAG. The schema is as follows:

Table A4.1: Contents of *jif.tsv*.

Variable	Type	Notes
journalid	numeric	
journalname	String	
jif	numeric	Journal impact factor. A journal's impact factor is a popular measure of its quality, calculated for year t as the number of times articles from years t-1 and t-2 were cited *by other articles* during year t, divided by the number of articles published during years t-1 and t-2.

In addition, we provide a new measure of journal impact: Journal Commercial Impact Factor (JCIF). Just like JIF is a journal-level measure of quality, it is possible to build a journal-level measure of appliedness or commercial relevance by replacing paper-to-paper citations by patent-to-paper citations. Bikard and Marx (2019) introduced this concept and calculated it for the Web of Science; here, we calculate JCIF for MAG. That paper should be cited if the JCIF data available here are used.

Table A4.2: Contents of *jcif.tsv*.

Variable	Type	Notes
journalid	numeric	
journalname	String	
jcif	numeric	Journal commercial impact factor. A journal's commercial impact factor is calculated for year t as the number of times articles from years t-1 and t-2 were cited *by patents* during year t, divided by the number of articles published during years t-1 and t-2.

Finally, we provide an aggregation of the more than 200,000 fields automatically extracted from the papers themselves. We mapped the MAG subjects to 6 OECD fields and 39 subfields, defined here: http://www.oecd.org/science/inno/38235147.pdf. Clarivate provides a crosswalk between the OECD classifications and Web of Science fields, so we include WoS fields as well. This file is magfield_oecd_wos_crosswalk.zip.

Table A4.2: Contents of *magfield_oecd_wos_crosswalk.tsv*.

Variable	Type	Notes
paperid	numeric	Unique identifier for each paper in the Microsoft Academic Graph.
paperfieldid	paperid, fieldid	PaperFieldsOfStudy.txt (1,2)
oecd_field	String	One of six top-level OECD fields.
oecd_subfield	String	One of 39 OECD subfields.
wosfield	String	One of 251 Web of Science fields.