Reliance on Science in Patenting

Matt Marx† & Aaron Fuegi††

3 June 2019

Abstract: Citations from patents to other patents have frequently been employed in studies of innovation, but these citations have many limitations. By contrast, citations from patents to *non*-patent materials—especially scientific articles—promise to be more useful but are much more difficult to discern given that they appear in patent documents as unstructured text. We present methods for automatically linking patents to scientific papers from 1800-2018 and share the results publicly. Moreover, we characterize the performance of our algorithms and present ROC curves so that researchers can select data according to their sensitivity to false positives vs. false negatives. Our hope is that publicly-available patent citations to science fuel research on innovation, knowledge diffusion, technology commercialization, and other topics.

† 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.

PRELIMINARY AND INCOMPLETE, latest at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331686

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 15MM patent citations to science. The patent-paper linkages, as well as selected metadata on the articles (whether cited or not), are publicly available for download at **https://linksplit.io/reliance_on_science**.

Patent citations to science (PCS) have occasionally been used by scholars to measure innovation and the flow of knowledge, but not as often as patent-to-patent citations. This disparity seems odd, given the limitations of patent-to-patent citations as well as reports that PCS more accurately represents knowledge flows, and may be explained in two ways.

First, PCS are difficult to work with given that they appear in patent records as *unstructured* text strings. By comparison, patents simply cite each other by number. Thus the analyst 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 perhaps 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 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 use by fellow researchers.

Naturally, the process of automatically constructing PCS from unstructured text is not error-free. We characterize the performance of our linkage algorithm first by building a set of "known good" references to assess false negatives and then inspecting a random sample of output to assess false positives. Recognizing that researchers may have different tolerances for Type I vs. Type II errors, we provide a confidence score for each match as well as receiver-operating characteristic (ROC) curves indicating the tradeoff between recall and precision at each confidence level. Based on third-party assessment, we estimate that our algorithm can capture nearly 90% of patent citations to science with an accuracy rate of 99%.

The paper is organized as follows. We begin by motivating the use of PCS as opposed to patent-to-patent citations, explain the difficulties of constructing PCS, 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. Fourth, we describe the resulting patent-to-paper linkages and how fellow researchers can access them. We conclude by sketching research avenues opened up by the broad availability of PCS.

MOTIVATION

Since Jaffe, Trajtenberg, and Henderson (1993), scholars have sought to measure the transfer of knowledge of academic discoveries to industry via patent citations. When filed, each patent must list prior art upon which it builds and from which the proposed invention is differentiated. Prior art may include either previously filed patents or "non-patent" information such as product

brochures, websites, government testimony, or scientific articles. As such, prior art citations from patents serve as a paper trail by which the flow of knowledge can be traced.

Citations from patents to other patents have been widely used by economists to study knowledge spillovers (e.g., Henderson, Jaffe, and Trajtenberg 1993; MacGarvie 2006; Agrawal, Cockburn, and McHale, 2006; Singh and Marx, 2013, among many others). One reason researchers have so readily adopted patent-to-patent citations in their research is that these are easily identified simply by matching patent numbers. In Figure 1, for example, it is straightforward to conclude that patent 6,789,062 cites patent 5,737,487.

Figure 1 about here

The ease of processing patent-to-patent citations is however be offset, at least in part, by their disadvantages. Many citations to patents are added by patent examiners and may not represent actual knowledge flows (Alcácer and Gittelman 2006; Alcácer, Gittelman, and Sampat 2009). Although scholars have made available machine-readable distinctions between examiner and applicant-supplied citations since Thompson's (2006) initial exploitation of the two citation types, even applicant-supplied patent-to-patent citations may not represent true knowledge flows. Lampe (2012) demonstrates that applicants may strategically cite—or not—other patents in order to facilitate the application's process through the patent examination procedure. As such, many patent-to-patent citations may be added by patent lawyers instead of inventors. For all of these reasons, patent-to-patent citations may serve at best as noisy indicators of knowledge flows.

These shortcomings are exacerbated when seeking to measure the flow of knowledge from universities to industry—including whether academic discoveries become commercialized. As Belenzon and Schankerman (2013) note, barely 10% of academic discoveries are patented. Even

yeoman efforts to track down commercialization outcomes for a population of university patents, as Shane (2001) did for nearly 1,400 patents assigned to MIT from 1980-1996, introduce selection bias as the primary vehicle for the dissemination of academic knowledge is publishing articles.

As noted above, patents cite not only other patents but also other sources including academic articles. Academic articles cited by patents may yield a more representative sample of academic discoveries than focusing on those discoveries that have been patented. These patent citations to science (PCS) not only promise to offer a view into a wider set of scientific discoveries; PCS may offer a clearer view as well.

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). 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.

Recognizing the promise of PCS for tracking knowledge flows, a small number of scholars have made the investment to use PCS in their research. Katila and Ahuja (2002) manually collect PCS for 124 industrial robotics firms. Watzinger, Treber, and Schnitzer (2018) assemble patent citations to scientific articles for 1,113 scientists at a German university. Belenzon and Schankerman (2013) measure the flow of knowledge from universities first using patent-to-patent citations and then using PCS for nearly 35,000 publications from 184 research universities. Arora, Belenzon, and Sheer (2017) collect PCS for 4,736 firms from 1980-2006. Azoulay, Graff Zivin,

and Sampat (2011) do so for 9,483 elite academic life scientists. b assemble PCS for 1,186 patents from 79 firms during 1995-2001. Bikard and Marx (2018) retrieve citations from patents to 316 "twin" scientific papers in academia from Scopus.

Challenges

Given the advantages of PCS over patent-to-patent citations, why have more scholars not used PCS in their research? One obstacle to their adoption is the difficulty of measuring such citations. Unlike patent-to-patent citations, which as shown in Figure 1 are easily captured by matching patent numbers, researchers wishing to use PCS face at least three hurdles:

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).

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 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 hundreds of trillions 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.

Table 1 about here

Prior efforts to assemble PCS at scale

These challenges may explain in part why researchers have been slow to adopt PCS, and why several PCS studies have focused on small samples where results can be checked by hand. That said, at least four research teams have undertaken larger-scale efforts to assemble PCS.

Li, Azoulay, and Sampat (2018) gather citations from approximately 1.3 million life-sciences patents (i.e., the Chemicals and Drugs/Medical categories) to any article indexed by PubMed by 2010. Their algorithm requires four of the following fields to match: first author, page, volume, publication year, journal, and initial title. False positives and false negatives are assessed with a hand-collected sample of approximately 300 known PCS.

Gaetani and Bergolis (2015) generate all PCS from 4.8 million USPTO patents 1974-2010 to 32 million articles captured by WOS since 1945. The nature of their algorithm is described by

Ahmadpoor and Jones (2017), who also employ their data, as matching on first author, publication year, volume, page number, and shared words in the journal name and paper title. Neither false-positive nor false-negative rates are reported.

Fleming et al. (2018) generate all PCS from USPTO patents 1947-2017 to 9.6 million articles captured by the Web of Science since 1945. They begin with an exact match for first author surname, journal name (including abbreviations), title, and publication year. On a second pass, they look for matches without the title but where the first author surname, publication year, volume, issue, and first page all match. They report no false positives in a random sample of 100 and 20% false negatives in a known-good sample of 169 references.

Similar to Fleming et al. (2018), Knaus & Palzenberger (2018) link patents to the Web of Science through 2017. They link not only patents from the USPTO but also the WIPO and EPO. The algorithm is described in detail, and the output matches are scored with a confidence level. Performance metrics are calculated at each confidence level.

Jefferson, et al. (2018) report linkages from 35 patent jurisdictions (1950-2015) to 9.8MM articles from PubMed and CrossRef. They use these matches to determine measures of industry influence for 200 leading global research institutions. Methods and the number of matches found are described, with matches limited to those with confidence score of 0.9; however, performance of the matching algorithm in terms of false-positives and false-negatives is not reported.

The linkages generated by Gaetani and Bergolis, Fleming et al., and Knaus & Palzenberger are not generally available to researchers. Because they used WOS as their list of scientific articles, their licensing agreement in all likelihood does not allow them to share the resulting data. Even if they were able to share raw counts of citations to WOS under their license—which

is unclear—it would be impossible for researchers to verify the accuracy of the resulting PCS. Linkages developed using open-source data such as PubMed and CrossRef can be made available, as Jefferson, et al. (2018) have done via query-based website access.[1]

The Microsoft Academic Graph

Most researchers studying the diffusion of knowledge have relied on patent-to-patent citations, which are more easily assembled than PCS. To date, researchers constructing PCS have used either PubMed or WOS as repositories of academic articles. PubMed is limited in scope by comparison to WOS, but the proprietary nature of WOS precludes posting derivative works publicly for replication, assessment, or use by other researchers.

WOS captures approximately 55 million articles whereas online repositories such as Google Scholar may have more than triple this number (Orduña-Malea, et al, 2014). Moreover, WOS contains few articles prior to 1945. Google Scholar's index would make an attractive candidate to use as a repository of academic articles, but it is not available for download. Further, Google has taken steps to prevent automated harvesting of its content.

The Microsoft Academic Graph (MAG) boasts a repository of more than 150 million publications since the year 1800 (Sinha et al, 2015), likely rivaling and possibly exceeding coverage by Google Scholar. As such, it promises to yield a larger set of PCS than would be found using the proprietary WOS or field-specific subsets such as PubMed. Furthermore, unlike Google Scholar, MAG is free to download and distributed under the Open Data Commons Attribution (ODC-By) license, which permits the creation and distribution of derivative works

[1] The data from Jefferson, et al. (2018) have been extended to include matches to the Microsoft data and are updated biweekly. At lens.org, one may search for citations either via PubMed IDs, DOIs, or patent numbers.

with acknowledgment. Thus it is possible to use MAG as the target set of academic articles for matching and publish the resulting dataset for evaluation and use by other researchers.

ESTABLISHING LINKAGES BETWEEN PATENTS AND SCIENTIFIC PAPERS

This section describes our algorithm for linking 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. From 1947-2018 we found 36,020,060 non-patent references from 3,095,844 patents.

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 Figure 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.

The remainder of this section describes the sequence of steps in our linking algorithm. 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[2] 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

[2] 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. The full set of exclusionary terms will be available when code is posted.

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. For example, if we started with only the following two unstructured lines:

> Wisenschmidt et al., "Developing a programmed restriction endonuclease for highly specific DNA cleavage," Nucleic Acids Res., 33(22):7039-47 (2005). cited by applicant.
>
> "lee, j. h. et al., ""immunnoassay of prostate-specific antigen (psa) using resonant frequency shift of piezoelectric nanomechanical microcantilever"", biosensors and bioelectronics, 20: 2157-2162 (2005). cited by other.

We would segment these into 34 files containing copies of one or both of these unstructured lines: Weisenschmidt.txt, Developing.txt, Programmed.txt…33.txt, 22.txt, 7039.txt…and so on. The file for cited.txt would contain a copy of both of these unstructured lines, whereas the others would contain just the unstructured line that contained that alphanumeric string.

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.

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. It is true that author names appear later in the unstructured string, but we elected not to apply flexible author-name matching to the entire strong for computational reasons. However, we find exact matches to the author anywhere in the unstructured string.

Often, the year is missing or misspecified. When misspecified, usually it is 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 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 page number. (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, but exact thresholds and matching terms will be available once the code is posted.

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) compose 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 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.

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). In the next step, we characterize the performance of matching at various confidence levels.

Performance Characterization

Our algorithm finds PCS linkages for 17,684,725 non-patent references in patents from 1947-2018. These represent PCS from 1,620,494 patents to 3,261,993 papers. However, an algorithm like ours will make both Type I and Type II errors. In this section, we characterize the performance of our algorithm.

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, we provide a large set of matches along with confidence scores, accompanied by the following characterization of precision vs. recall at each confidence level.

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. A research assistant checked these independently, and then reviewed the results with one of the authors. The number found to be incorrect at each confidence level are listed in the third column of Table 2. The corresponding 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 2 about here

As is visible from Table 2, 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. Most of the errors found, especially at higher confidence levels, involve similar papers by the same author.

To check recall (false negatives), a test set of "known good" references must be created. Of course, the algorithm developers cannot involve themselves either in the evaluation of output and especially creation of the known-good references, lest the algorithm be overfitted to these test cases. Accordingly, we tasked multiple research assistants with creating the known-good cases from a random sample of 1000 unstructured non-patent references. 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 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 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 3 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 3 about here

Researchers may have different preferences for recall vs. precision. The receiver-operator curve (ROC) in Figure 2 plots recall against false negatives (i.e. one minus precision), using the recall statistics from Table 3.

Figure 2 about here

Available Output

Researchers and the general public are invited to download both the PCS linkages as well as selected MAG metadata. These are available at the Inter-university Consortium for Political and Social Research (ICPSR) at the **https://linksplit.io/reliance_on_science** with this paper as 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, "other", or unknown and is given a confidence score from 1-10. The schema for this output file is detailed in Appendix 1. Researchers interested only in the number of PCS per patent may find this file sufficient; however, we suspect that most researchers will want to know about papers that were referenced, as well as papers not referenced. 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. These files and schemas are described in Appendix 2.

Appendices 1 and 2 about here

Known Limitations and Future Directions

There are three 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. Second, if the author's name is misspelled but does not occur in the first four words of the unstructured string, our algorithm will not find the match. A third category is references that a) omit both the longest and second-longest 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.

In addition, the 36MM patent references provided by the USPTO are likely limited to "front page matter" and do not include references embedded within the narrative of the patent text. Bryan and Ozcan (2018) report linkages from 132,872 PubMed articles to the narrative text of patents from 2005-2015. We see incorporating in-text citations is an important next step.

CONCLUSION

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 this section, we sketch several research questions that are enabled by these data.

Knowledge diffusion

Perhaps the most immediately-obvious application of PCS is the study of knowledge diffusion. Since Jaffe et al. (1993), several scholars have sought to understand the flow of knowledge using citations from patents to other patents. The limitations of patent-to-patent citations, including that many of them are added by patent examiners or attorneys instead of the inventors themselves, are ameliorated somewhat. At least in the USPTO jurisdiction, references to science are rarely added by examiners and moreover are more often supplied by the inventors themselves than their attorneys (Bikard & Marx, 2018). Belenzon & Schankerman (2013) as well as Azoulay, Graff Zivin, and Sampat (2011) have already leveraged PCS to study knowledge diffusion in targeted subsamples of selected universities and elite life scientists.

Firms, open innovation, and reliance on academic science

Firms decreasingly invest in internal research and development (Arora, Belenzon, and Patacconi, 2018). How, then, do they innovate? Appropriable science that has been published is a key source of potential innovations, but to which scientific discoveries do firms pay attention? Again, PCS have been used to address this question for particular industries (Katila & Ahuja, 2002; Arora, et al., 2018) and for simultaneous discoveries (Bikard, 2018; Bikard & Marx, 2018), but the same can now be investigated at scale.

Cross-community networks of invention

Economists and sociologists alike have published myriad studies of teams, including scientific teams. To date, however, these studies have largely occurred within the community of those publishing papers (i.e., coauthors) or those applying for patents (i.e., co-inventors). Do these networks intersect? If so, how and when, and does it matter? Patent citations to science enable researchers to check for overlap between authors on a paper and inventors on a patent, thus creating cross-community networks.

Impact and commercialization of academic science

Scholars and policymakers alike have long sought to understand the impact of multibillion-dollar annual investments in academic research. Documenting technology transfer is challenging, as many university inventions are not patented and databases of university licensing can be difficult to obtain. Although not affording a complete picture, PCS can help to inform our understanding of how academic discoveries are commercialized by enabling researchers to link patents to the science they depended on. Marx & Hsu (2018) do so for a subset of 20,000 duplicate discoveries in academic science, using the overlap between authors on a paper and inventors on a patent to establish paper-patent pairs (Murray, 2002) assigned to startup companies. Such efforts can be expanded to include estimates of the impact of university research on productivity, entrepreneurship, employment, etc.

Even when scholars are not directly studying the above topics, they may find PCS useful in assembling critical controls or categorizations. As one example, scholars often wish to control for the reliance of a patent—or a firm, or an industry—on scientific input. Although researchers have often used counts of non-patent references as a proxy for scientific heritage (e.g., Sorenson

& Fleming, 2004), as Cassiman et al. (2008) caution, many non-patent references are not to scientific articles but rather product brochures, patent office actions, and other artifacts having little to do with science. Now that non-patent references have been mapped directly to papers, researchers can include counts of PCS with greater confidence.

This is only a small sampling of possible topics to be addressed. Our hope is that this dataset will enable researchers to tackle topics that were previously infeasible or cost-prohibitive.

REFERENCES

Agrawal, Ajay, Iain Cockburn, and John McHale. "Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships." Journal of Economic Geography 6.5 (2006): 571-591.

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

Alcácer, Juan, and Michelle Gittelman. 2006. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations." Review of Economics and Statistics 88 (4): 774–79.

Alcácer, Juan, Michelle Gittelman, and Bhaven Sampat. 2009. "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis." Research Policy 38 (2): 415–27.

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.

Arora, Ashish, Sharon Belenzon, and Lisa Sheer. "Back to Basics: Why do firms invest in research?" Mimeo.

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, Michaël. "Made in Academia: The Effect of Institutional Origin on Inventors' Attention to Science." Organization Science (2018).

Bikard, Michael and Matt Marx. "Hubs as lampposts: Academic location and firms' attention to science." Mimeo.

Bryan, Kevin and Yasin Ozcan (2018). "The Impact of Open Access Mandates on Invention". Mimeo.

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.”

Gaetani, Ruben, and M. Li Bergolis. "The economic effects of scientific shocks." Unpublished Manuscript (2015).

Jaffe, Adam B. "Real effects of academic research." The American Economic Review (1989): 957-970.

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. 1993. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” The Quarterly Journal of Economics 108 (3): 577–98.

Jaffe, A. and M. Trajtenberg (eds.) Patents, Citations and Innovations. Cambridge, MA: The MIT Press, 2002.

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.

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.

Lampe, Ryan. 2012. “Strategic Citation.” Review of Economics and Statistics 94 (1): 320–33.

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.

MacGarvie, Megan. 2006. “Do Firms Learn from International Trade?” Review of Economics and Statistics 88 (1): 46–60.

Marx, M. and D. Hsu. 2018. “The Entrepreneurial Commercialization of Science: Evidence from “Twin” Discoveries.”

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. Research policy, 31(8-9), 1389-1403.

Orduña-Malea, Enrique, Juan Manuel Ayllón, Alberto Martín-Martín, Emilio Delgado López-Cózar. “About the size of Google Scholar: Playing the numbers.” Mimeo, 2014.

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.

J. Singh and M. Marx. “Geographic Constraints on Knowledge Diffusion: Political Borders vs. Spatial Proximity.” Management Science 59(9):2056-2078 (2013).

S. Shane (2001). “Technological opportunities and new firm creation,” *Management Science*, 47(2): 205-220.

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.

Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. Research policy, 33(10), 1615-1634.

Thompson, Peter. "Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations." The Review of Economics and Statistics 88.2 (2006): 383-388.

Watzinger, Martin, Lukas Treber, and Monika Schnitzer. “Universites 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. Each non-patent reference is preceded by the number of the patent making the citation.

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," Nucleic 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

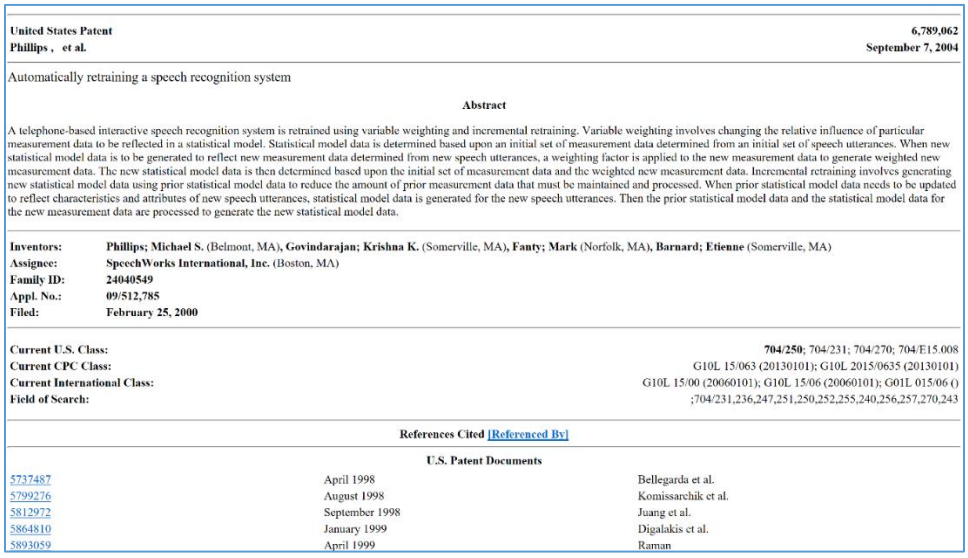
Table 2: Precision (1 minus false positives) in a random sample of 100 per confidence level

(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 3: Recall (1 minus 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: Header and patent-to-patent citations for patent 6,789,062.



United States Patent **6,789,062**
Phillips , et al. **September 7, 2004**

Automatically retraining a speech recognition system

Abstract

A telephone-based interactive speech recognition system is retrained using variable weighting and incremental retraining. Variable weighting involves changing the relative influence of particular measurement data to be reflected in a statistical model. Statistical model data is determined based upon an initial set of measurement data determined from an initial set of speech utterances. When new statistical model data is to be generated to reflect new measurement data determined from new speech utterances, a weighting factor is applied to the new measurement data to generate weighted new measurement data. The new statistical model data is then determined based upon the initial set of measurement data and the weighted new measurement data. Incremental retraining involves generating new statistical model data using prior statistical model data to reduce the amount of prior measurement data that must be maintained and processed. When prior statistical model data needs to be updated to reflect characteristics and attributes of new speech utterances, statistical model data is generated for the new speech utterances. Then the prior statistical model data and the statistical model data for the new measurement data are processed to generate the new statistical model data.

Inventors: **Phillips; Michael S.** (Belmont, MA), **Govindarajan; Krishna K.** (Somerville, MA), **Fanty; Mark** (Norfolk, MA), **Barnard; Etienne** (Somerville, MA)
Assignee: **SpeechWorks International, Inc.** (Boston, MA)
Family ID: **24040549**
Appl. No.: **09/512,785**
Filed: **February 25, 2000**

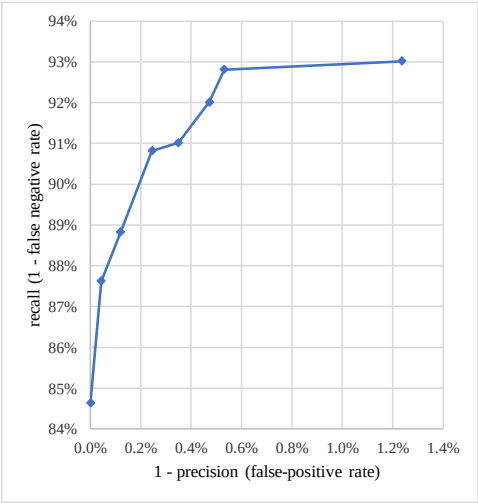
Current U.S. Class: **704/250**; 704/231; 704/270; 704/E15.008
Current CPC Class: G10L 15/063 (20130101); G10L 2015/0635 (20130101)
Current International Class: G10L 15/00 (20060101); G10L 15/06 (20060101); G01L 015/06 ()
Field of Search: ;704/231,236,247,251,250,252,255,240,256,257,270,243

References Cited [Referenced By]

U.S. Patent Documents

5737487	April 1998	Bellegarda et al.
5799276	August 1998	Komissarchik et al.
5812972	September 1998	Juang et al.
5864810	January 1999	Digalakis et al.
5893059	April 1999	Raman

Figure 2: ROC curve for PCS linkages



Appendix 1: Schema for patent citations to science (PCS) output files

The main output file, available at **https://linksplit.io/reliance_on_science**, 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 Oth=from other Unk = if unspecified in the unstructured reference
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 likely very few correct matches at those levels. Even at confidence score 3, about half of the matches are incorrect. Most users will want to use matches only with a score of 4 or higher.

Appendix 2: Files for Microsoft Academic Graph metadata

Also available at **https://linksplit.io/reliance_on_science**, in the subdirectory *<selected MAG metadata 2018>,* 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. The code used to reduce the original MAG fields to the files provided, *create_subsets_tsv_nochop.sh,* is located in the subdirectory <original MAG documentation and translation> of <selected MAG metadata 2018>, along with the original MAG documentation, *Microsoft-Academic-ADLS-ReadMe-end-2018.pdf*.

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	Journals.txt (1,3)	